Statistics in Sociology, 1950-2000: A Selective Review

Adrian E. Raftery

Sociological Methodology, Vol. 31. (2001), pp. 1-45.

Stable URL:
http://links.jstor.org/sici?sici=0081-1750%282001%2931%3C1%3ASIS1AS%3E2.0.CO%3B2-Q

Sociological Methodology is currently published by American Sociological Association.

1

STATISTICS IN SOCIOLOGY, 1950–2000: A SELECTIVE REVIEW

*Adrian E. Raftery**

Statistical methods have had a successful half-century in sociology, contributing to a greatly improved standard of scientific rigor in the discipline. I identify three overlapping postwar generations of statistical methods in sociology, based on the kinds of data they address. The first generation, which started in the late 1940s, deals with cross-tabulations and focuses on measures of association and log-linear models, perhaps the area of statistics to which sociology has contributed the most. The second generation, which began in the 1960s, deals with unit-level survey data and focuses on LISREL-type causal models and event-history analysis. The third generation, starting to emerge in the late 1980s, deals with data that do not fall easily into either of these categories, either because they have a different form, such as texts or narratives, or because dependence is a crucial aspect, as with spatial or social network data. There are many new challenges, and the area is ripe for statistical research; several major institutions have recently launched new initiatives in statistics and the social sciences.

1. INTRODUCTION

To mark the year 2000, the *Journal of the American Statistical Association* published a series of about 50 short vignettes, each about some aspect of statistical development in the century that was ending. The idea was to

I am very grateful to Mark Becker, Mark Handcock, Don Rubin, Tom Snijders, Rob Warren, Yu Xie, and Kazuo Yamaguchi for extremely helpful comments that greatly improved the manuscript.

*University of Washington

summarize some of the best work and to highlight potentially fruitful areas of future research. I wrote the vignette about statistics in sociology (Raftery 2000). Other vignettes of possible interest to *Sociological Methodology* readers include the ones on contingency tables and log-linear models (Fienberg 2000), causal inference in the social sciences (Sobel 2000), demography (Xie 2000), political methodology (Beck 2000), psychometrics (Browne 2000), and empirical methods in legal science (Eisenberg 2000).

Many colleagues sent me comments on the first draft, quite a few of them correctly pointing out important developments that I had missed. It was impossible to rectify this given the small amount of space allocated by JASA, but *Sociological Methodology* editors Michael Sobel and Mark Becker invited me to submit an expanded version that would provide more appropriate coverage of this dynamic field.

The roots of sociology go back to the mid-nineteenth century and to seminal work by Auguste Comte (who invented the word "sociology"), Karl Marx, Max Weber, and Emile Durkheim on the kind of society then newly emerging from the Industrial Revolution. Sociology has used quantitative methods and data from the beginning. Comte, who launched the discipline, was quite explicit about its grounding in statistical data. Durkheim's (1897) *Le Suicide*, for example, made extensive use of statistical data.

However, prior to World War II, the data tended to be fragmentary, often bordering on the anecdotal, and the statistical methods simple and descriptive. Camic and Wilson (1994) identified Franklin H. Giddings as the father of quantitative sociology in America. Giddings, who was appointed professor of sociology at Columbia in 1894 and died in 1931, defined sociology as a field that studies social phenomena at the aggregate level. He held that statistical analysis in sociology consists largely of counting the individuals in each of several categories and finding average characteristics of each category. From a modern statistical perspective, a striking feature of his work was his relative lack of concern with variation.

Since then, the data available have grown in complexity, and statistical methods have been developed to deal with them. Much of this statistical development has been due to sociologists rather than statisticians; Clogg (1992) and the discussants of his article made this point emphatically, and documented it well. This partly reflects the fact that the number of statisticians working on sociological problems has always been relatively small. Statisticians have tended to work in greater numbers on

problems emerging from medicine, engineering and the biological sciences; this probably reflects the balance of available funding in the latter half of the twentieth century. There are some signs recently that this situation is changing, which I will mention at the end of the article.

The overall trend in sociology in the past 50 years has been toward more rigorous formulation of hypotheses, larger and more detailed data sets, statistical models growing in complexity to match the data, and a higher level of statistical analysis in the major sociological journals. Statistical methods have had a successful half-century in sociology, contributing to a greatly improved standard of scientific rigor in the discipline.

Sociology has made extensive use of a wide variety of statistical methods and models. I will focus here on the ones developed by sociologists, those whose development was directly motivated by sociological problems, and those that were first published in sociological journals. Many other methods, such as those for limited dependent variables (logistic regression springs to mind), have been used extensively in sociology, but they were primarily developed in other disciplines in response to other problems. Important though they are for sociology, I will mention these areas more briefly.

A major omission from this article is any in-depth discussion of statistical methods that have come to sociology from econometrics rather than from statistics; this would merit a separate review article in its own right. Econometrics has been very influential in sociological methodology, some would argue as much or more as statistics itself, but here I do not review this important influence except incidentally.

At the risk of controversy, I will classify statistical methods in sociology by the kind of data that they address, rather than by the method itself. I will distinguish three postwar generations of statistical methods in sociology, each defined by the kind of data to which they are most often applied: (1) cross-tabulations, (2) unit-level survey data, and (3) newer data forms. Like real generations, these intellectual generations overlap and the boundaries between them are not clear-cut; they all remain active today, albeit at different levels of maturity, and even their starting points are not uniquely defined.

In the period starting after World War II, much of the data that sociologists had to work with came in the form of cross-tabulations of counts from surveys and censuses. The first generation of methods I will discuss deals with data of this kind. Typically these cross-classifications involved only a small number of variables such as sex, age group and

occupational category; social mobility tables provided the canonical example for much of the methodological work. This is perhaps the area of statistics to which sociologists have contributed the most; indeed, it could be argued that sociologists have dominated this subfield and that the methods they have developed have been diffusing out from sociology into other disciplines. Schuessler (1980) is a survey that largely reflects this first-generation work.

By the early 1960s, sociologists no longer had to rely on cross-tabulations of counts, and unit-level data from surveys that measured many variables were becoming available. Computing power was also developing to the point where it could handle such data fairly easily. The second generation of methods was developed to deal with data of this kind. This generation of methods was galvanized by Blau and Duncan's (1967) highly influential book, *The American Occupational Structure*, and also by the establishment of *Sociological Methodology* in 1969, and that of *Sociological Methods and Research* in 1972 as publication outlets. Edgar Borgatta established both of these publications, the second when it became rapidly apparent that there was both the supply and demand for more articles than could be published by *Sociological Methodology* alone. These developments marked the coming of age of research on quantitative methodology in sociology.

By the late 1980s, sociologists had conceived the ambition of analyzing data that do not fit easily into the standard straitjackets of cross-tabulations or data matrices (although they can sometimes be forced into it). These include texts or narratives, and data in which dependence is a crucial aspect, such as social network data and data in which spatial referencing is a crucial aspect. They also include data sets that combine multiple types of data, such as satellite images, ethnographic accounts, and quantitative measurements. The third generation of methods is being developed to address data such as these. As befits its youth, so far it is a lively and exciting grab bag of ideas and developments, not having yet achieved the well-organized maturity of the first two generations.

My classification of statistical methods in sociology into generations defined by the kind of data addressed, rather than the kind of method used, does not reflect the usual organization of graduate training, and it is bound to be somewhat controversial. Perhaps for reasons of convenience and efficiency in training, the major methods of sociology have tended to be grouped together under categories such as regression models, limited dependent variable models, log-linear models, structural equation mod-

els, event-history analysis, and so on. However, I have found it easier to attempt to discern past trends and to think about future developments by focusing on the types of data that motivate the development of the methods in the first place.

We have come a long way in the past 50 years. Today, much sociological research is based on the reanalysis of large high-quality survey sample data sets, often collected with public funds and publicly available to researchers, with typical sample sizes in the range 5,000 to 20,000, or greater. This has opened the way to easy replication of results and has helped to produce standards of scientific rigor in sociology comparable to and greater than those in many of the natural and medical sciences. Perhaps in part because of this, social statistics has recently started a rapid expansion as a research area, and several major institutions have launched initiatives in this area in the past few years.

2. THE FIRST GENERATION: CROSS-TABULATIONS

2.1. *Categorical Data Analysis*

Initially, much of the data that quantitative sociologists had to work with came in the form of cross-classified tables, and so it is not surprising that this is perhaps the area of statistics to which sociology has contributed the most. A canonical example has been the analysis of social mobility tables, usually in the form of two-way tables of father's against respondent's occupational category; typically the number of categories used is between 5 and 17.

At first the focus was on measures of association, or mobility indices as they were called in the social mobility context (Glass 1954; Rogoff 1953), but these indices failed to do the job of separating structural mobility from exchange (or circulation) mobility. The solution to this key problem in the analysis of mobility tables turned out to require explicit probability models for the tables. Birch (1963) proposed the log-linear model for the observed counts $\{x_{ij}\}$, given by

$$\log(E[x_{ij}]) = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}, \tag{1}$$

where i indexes rows and j columns, $u_{1(i)}$ and $u_{2(j)}$ are the main effects for the rows and columns, and $u_{12(ij)}$ is the interaction term, measuring departures from independence. This provided the overall framework needed for

the rigorous analysis of mobility and similar tables. However, the difficulty with model (1) in its original form for social mobility and similar tables is that the number of parameters is too large for inference and interpretation. For example, in the U.S. data sets 17 categories were used, so the interaction term involves $16^2 = 256$ parameters.

To make progress, it was necessary to model the interaction term parsimoniously (i.e., with few parameters), but in a way that fits the data. A successful general approach to doing this is the association model of Duncan (1979) and Goodman (1979):

$$u_{12(ij)} = \sum_{m=1}^{M} \gamma_m \alpha_i^{(m)} \beta_j^{(m)} + \phi_i \delta(i,j), \tag{2}$$

where $\delta(i,j) = 1$ if $i = j$ and 0 otherwise. In (2), $\alpha_i^{(m)}$ is the score for the ith row on the kth scoring dimension, and $\beta_j^{(m)}$ is the corresponding score for the jth column; these can be either specified in advance or estimated from the data. The last term allows a different strength of association on the diagonal. (The model [2] is unidentified as written; various identifying constraints are possible.) This is often called the RC(M) model. In most applications to date, $M = 1$; the first genuine substantive application of the model in sociology with $M > 1$ was to labor market experiences and outcomes by Clogg, Eliason, and Wahl (1990).

Goodman (1979) initially derived this model as a way of describing association in terms of local odds ratios. Goodman (1985) has shown that this model is closely related to canonical correlations and to correspondence analysis (Benzécri 1976), and provides an inferential framework for these methodologies. When the categories are ordered, the uniform association model with $\alpha_i = \beta_i = i$ is a useful starting point (Haberman 1979). In this model, the odds ratios in all 2×2 subtables are equal, so this can be viewed as a discrete analog to the bivariate normal distribution, with $\gamma \equiv \gamma_k$ specifying the correlation.

Table 1 shows the actual counts for a reduced version of the most extensive U.S. social mobility study, and the fitted values from an association model; the model accounts for 99.6 percent of the association in the table and its success is evident. Hout (1984) extended the range of application of these models by modeling the scores and diagonal terms in (2) as sums or products of covariates, such as characteristics of the occupational categories in question; this is an extension of Birch's (1965) linear-by-linear interaction model.

TABLE 1
Observed Counts from the Largest U.S. Social Mobility Study and Expected Values from a Goodman Association Model with Four Degrees of Freedom.*

Father's Occupation	Son's Occupation									
	Upper Nonmanual		Lower Nonmanual		Upper Manual		Lower Manual		Farm	
	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.
Upper nonmanual	1414	*1414*	521	*534*	302	*278*	643	*652*	40	*42*
Lower nonmanual	724	*716*	524	*524*	254	*272*	703	*698*	48	*43*
Upper manual	798	*790*	648	*662*	856	*856*	1676	*1666*	108	*112*
Lower manual	756	*794*	914	*835*	771	*813*	3325	*3325*	237	*236*
Farm	409	*386*	357	*409*	441	*405*	1611	*1617*	1832	*1832*

*Sample size is 19,912.
Source: Hout (1983).

This methodology has also made it feasible to model relatively high-dimensional tables with large numbers of categories in a parsimonious and interpretable way. This has led to important discoveries, including Hout's (1988) finding that social mobility has been increasing in the United States. This is a subtle result because of the complex nature of the data underlying it, and it would have been hard to discover it without using the association model methodology. This substantive result was confirmed and refined in Ganzeboom, Luijkx, and Treiman's (1989) discovery, based on several hundred social mobility tables from different countries at different time points, that social mobility has been increasing by about 1 percent a year in industrialized countries in the second half of the twentieth century.

Biblarz and Raftery (1993) and Biblarz, Raftery and Bucur (1997) adapted the models to higher dimensional tables to study social mobility in nonintact families. The tables they used had up to about 7,000 cells and five dimensions: (1) father's occupation, (2) offspring's occupation, (3) gender, (4) race, and (5) period. Thus standard log-linear models would not have revealed anything, but association modeling, extending the models mentioned earlier, did provide interpretable results, parameter estimates, and conclusions. They showed that occupational resemblance is weaker in nonintact families than in intact ones, that offspring raised by working single mothers succeed much better on average than those from other nonintact families, and that these patterns have remained essentially

constant from the 1960s through the 1990s, in spite of the many changes in family structure and occupational distribution as well as the relationship between gender, race, and occupational and labor force status. Other important applications of log-linear and related models include the analysis of sex segregation (Charles and Grusky 1995), and assortative mating (Kalmijn 1991). From sociology, the use of association models has diffused to other disciplines, such as epidemiology (Becker 1989).

One common reason for analyzing tables with more than two dimensions is to assess how two-way associations vary across a third (or several other) dimension(s). Yamaguchi (1987) and Xie (1992) have proposed specific forms of the higher-dimensional association model that are adapted for this purpose, and these were unified and extended by Goodman and Hout (1998). A particularly appealing aspect of the latter approach is the availability of a range of graphical displays that facilitate the interpretation of the rather complex data and model parameters that arise in this setting.

These models are for situations with discrete independent variables. Perhaps the most successful models for the dependence of cross-classifications on *continuous* independent variables are Sobel's (1981, 1985) diagonal mobility models. These have been applied in a variety of settings, for example to marital fertility (Sorensen 1989), cultural consumption (De Graaf 1991), and voting behavior (Weakliem 1992).

An intuitive alternative formulation of the basic ideas underlying (1) and (2) is in terms of marginal *distributions* rather than the main effects in (1). The resulting marginal models specify a model for the marginal distributions and a model for the odds ratios, and this implies a model for the joint distribution that is not log linear (Lang and Agresti 1994; Becker 1994; Becker and Yang 1998). The first substantive application of these models in sociology was to modeling social mobility (Sobel, Becker, and Minick 1998).

2.2. *Latent Class Models*

An alternative approach that answers different questions is the latent class model (Lazarsfeld 1950; Lazarsfeld and Henry 1968; Goodman 1974a,b). In its basic form, this represents the distribution of counts as a finite mixture of distributions within each of which the different variables are independent. The model was introduced to account for observed associations in multivariate discrete data, the original motivation being somewhat akin to that for factor analysis for multivariate continuous data.

Hagenaars (1988, 1990) has extended the latent class model to the situation where each component in the mixture can exhibit dependence. Clogg (1995) gives a survey of this area. There have been many applications of this model. One interesting recent application to criminology is by Roeder, Lynch, and Nagin (1999).

This basic model has been formulated and used in other contexts. Chickering and Heckerman (1997) formulated it as a Bayesian graphical model with one hidden node. This formulation facilitates estimation of latent class models with many variables, and also makes it easier to estimate the model when there are missing data for some individuals, and to make inference about the missing data. Celeux and Govaert (1991) used the same basic model for clustering multivariate discrete observations, again potentially with large numbers of variables.

2.3. *Hypothesis Testing and Model Selection*

Sociologists often have sample sizes in the thousands, and so they come up early and hard against the problem that standard P-values can indicate rejection of null hypotheses in large samples, even when the null model seems reasonable theoretically and inspection of the data fails to reveal any striking discrepancies with it. The problem is compounded by the fact that there are often many models rather than just the two envisaged by significance tests, and by the need to use stepwise or other multiple comparison methods for model selection (e.g., Goodman 1971). By the early 1980s, some sociologists were dealing with this problem by ignoring the results of P-value-based tests when they seemed counterintuitive and by basing model selection instead on theoretical considerations and informal assessment of discrepancies between model and data (e.g., Fienberg and Mason 1979; Hout 1983, 1984; Grusky and Hauser 1984).

It was soon pointed out that this problem could be alleviated by basing model selection instead on Bayes factors (Raftery 1986a), and that this could be simply approximated for log-linear models by preferring a model if BIC—defined by BIC = Deviance − (Degrees of freedom) $\log(n)$—is smaller (Schwarz 1978; Raftery 1986b). For nested hypotheses, this can be viewed as defining a significance level for a test that decreases automatically with sample size. Since then, this approach has been used in many sociological applications of log-linear models. Kass and Wasserman (1995) showed that the approximation is quite accurate if the Bayesian prior used for the model parameters is a unit information

prior—i.e. a prior distribution that contains about the same amount of information as a single "typical" observation. Raftery (1995) indicated how the methodology can be extended to a range of other models.

Weakliem (1999) criticized the use of BIC on the grounds that the unit information prior to which it corresponds may be too diffuse in practice, leading BIC to tend to favor the null hypothesis too often. However, Raftery (1999) pointed out that the unit information prior does provide a reasonable representation of the prior knowledge of investigators who have some advance information, but not a great deal, about the parameter values for the model they are estimating. It can thus be viewed as approximating the situation where there is little prior information. A more knowledgeable investigator would have a tighter prior distribution, and thus might have a basis for rejecting a null hypothesis when BIC does not, but this would be based on prior information rather than on data, and this should be made explicit in any report that does so. BIC provides a conservative assessment of evidence: one can be quite confident of the reality of any "effect," evidence of whose existence is favored strongly by BIC. Weakliem's arguments can be viewed as implying that if real prior information is indeed available it should be used, and I would agree with this. This points toward using Bayes factors based on priors that reflect the actual information available; this is easy to do for log-linear and other generalized linear models (Raftery 1996).

3. THE SECOND GENERATION: UNIT-LEVEL SURVEY DATA

The second generation of statistical models responded to the availability of unit-level survey data in the form of large data matrices of independent cases. The methods that have proved successful for answering questions about such data have mostly been based on the linear regression model and its extensions to path models, structural equation models, generalized linear models and event-history models. For questions about the *distribution* of variables rather about than their predicted value, however, nonparametric methods have proved useful (Morris, Bernhardt, and Handcock 1994; Bernhardt, Morris, and Handcock 1995; Handcock and Morris 1998, 1999). We start by reviewing the development of the measurement of occupational status, which provided a major impetus for the growth of the second generation of methods.

3.1. *Measuring Occupational Status*

Occupational status is an important concept in sociology, and developing a useful continuous measure of it was a signal achievement of the field. It was important for the development of statistical methods in sociology because, starting in the early 1960s, it encouraged greater use of regression analysis and related methods among scholars with an interest in the sources and consequences of job-holding. These methodological approaches diffused rapidly into other areas of the discipline.

Initially, the status of an occupation was equated with its perceived prestige, as measured in national surveys beginning in the 1940s. However, surveys could measure the prestige of only a small number of the several hundred occupations identified in each decennial Census classification. To fill in missing prestige scores for the 1960 Census classification, Duncan (1961) regressed the prestige scores for the 45 occupations for which they were available on measures of the proportion of occupational incumbents who had completed high school and the proportion of incumbents who earned more than $10,000. He found that the predictions were very good ($R^2 = 0.91$) and that the two predictors were about equally weighted. Based on this, he created a predicted prestige score for all occupations in the 1960 classification, which became known as the Duncan Socioeconomic Index (SEI); the SEI later turned out to be a better predictor of various social outcomes than the prestige scores themselves. Duncan's initial work has been updated several times for subsequent Census classifications (Featherman and Stevens 1982; Nakao and Treas 1994; Hauser and Warren 1997), but it has recently been critiqued on conceptual and empirical grounds (Hauser and Warren 1997; Warren, Sheridan, and Hauser 1998).

In much social research, particularly in economics, current income is used as a predictor of social outcomes, but there are good reasons to prefer occupational status. It has proved to be a good predictor of many social outcomes. Jobs and occupations can be measured accurately, in contrast to income or wealth, whose measurement is plagued by problems of refusal, recall, and reliability. Also, occupational status is more stable over time than income, both within careers and between generations. This suggests that occupational status may actually be a better indicator of long-term or permanent income than (current) income itself. The status of occupations tends to be fairly constant both in time and across countries (Treiman 1977).

3.2. *The Many Uses of Structural Equation Models*

Figure 1 shows the basic path model of occupational attainment at the heart of Blau and Duncan (1967); see Duncan (1966). Wright (1921) introduced path analysis, and Blalock (1961) gave it a causal interpretation in a social science context. One of the important uses and motivations of structural equation models was to decompose a total effect into direct and indirect effects. Alwin and Hauser (1975) played an important role in showing how to do this for sociological data. See Freedman (1987) and Sobel (1998) for critiques, and Section 3.8 below for more discussion of causality in the social sciences.

Often, variables of interest in a causal model are not observed directly, but other variables are observed that can be viewed as measurements of the variables, or "constructs" of interest, such as prejudice, alienation, conservatism, self-esteem, discrimination, motivation or ability. Jöreskog (1973) dealt with this by maximum likelihood estimation of a structural equation model with latent variables; this is sometimes called a

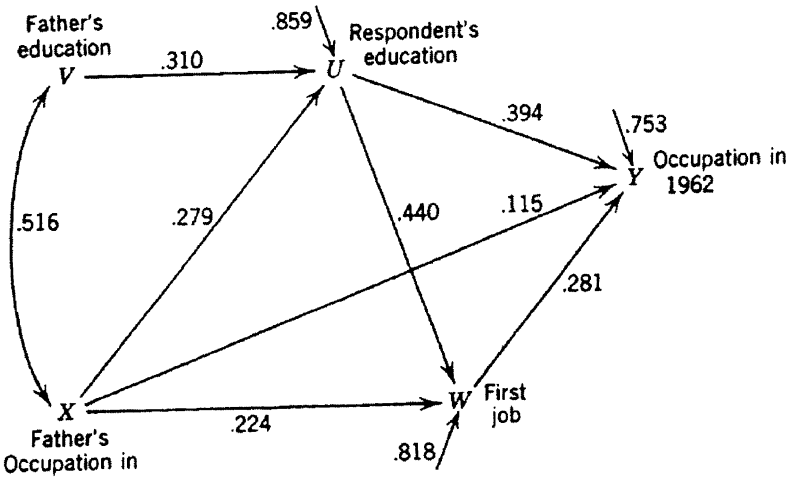

FIGURE 1. A Famous Path Model: The Process of Stratification, U.S. 1962. The numbers on the arrows from one variable to another are regression coefficients, 0.516 is the correlation between V and X, and the numbers on the arrows with no sources are residual standard deviations. All the variables have been centered and scaled.
Source: Blau and Duncan (1967).

LISREL model, from the name of Joreskog's software. Duncan (1975) played a big role in introducing these ideas into sociology, and Long (1984a,b) and Bollen (1989) provide well-written and accessible accounts geared to sociologists. A typical model of this kind is shown in Figure 2; the goal of the analysis is testing and estimating the strength of the relationship between the unobserved latent variables represented by the thick arrow. Diagrams such as Figures 1 and 2 have proved useful to sociologists for specifying theories and hypotheses and for building causal models.

The LISREL framework has been extended and used ingeniously for purposes beyond those for which it was originally intended. Muthén (1983) extended it to categorical variables, and Muthén (1997) showed how it can be used to represent longitudinal data, growth curve models, and multilevel data. Kuo and Hauser (1996) used data on siblings to control for unobserved family effects on socioeconomic outcomes, and cast the resulting random effects model in a LISREL framework.

The advent of graphical Markov models (Spiegelhalter et al. 1993), specified by conditional independencies rather than by regression-like relationships, is important for the analysis of multivariate dependencies, although they can seem less interpretable to sociologists. They have been particularly useful for propagating information about some variables through a system of dependent variables, to yield information about other, unobserved variables, as is needed, for example, in the construction of expert systems for medical diagnosis and other applications. They have been less used so far for inference and modeling in social research, per-

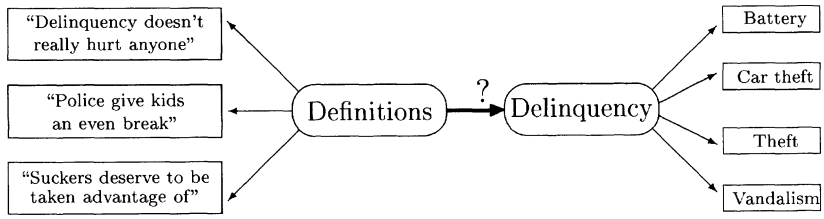

FIGURE 2. Part of a structural equation model to assess the hypothesis that learned definitions of delinquency cause delinquent behavior. The key goal is testing and estimating the relationship represented by the thick arrow. The constructs of interest, "Definitions" and "Delinquency," are not measured directly. The variables inside the rectangles are measured.
Source: Matsueda and Heimer (1987).

haps because sociological hypotheses tend to be formulated more often in terms of regression or causal relationships than in terms of conditional independencies between variables.

The relationship between graphical Markov models and structural equation models has begun to be understood (Koster 1996; Spirtes et al. 1998). Also, the LISREL model seems ideally suited to Gibbs sampling and Markov chain Monte Carlo (MCMC) methods (Gilks et al. 1996), and this is likely to permit useful extensions of the framework (Raftery 1991; Arminger 1998; Scheines, Hoijtink, and Boomsma 1999).

3.3. *Event-History Analysis*

Unit-level survey data often include or allow the reconstruction of life histories. These include the times of crucial events such as marriages, divorces, births, commitals to and releases from prison, job changes, or going on or off welfare.

Prior to 1972, two approaches were available for the analysis of the distribution of the time to a single event such as death, and of the factors influencing it. One was life table analysis from demography, but this did not allow easy analysis of the factors influencing time to an event. The other was regression analysis of the observed times to the event, but this was plagued by censoring, and by the often extreme nonnormality of the response.

This field was revolutionized by the introduction of the Cox (1972) proportional hazards model, which brought together these two approaches. Tuma (1976) and Tuma and Hannan (1984) generalized this approach to allow for repeated events, for multiple types of events, such as marriages and divorces, and for events consisting of movement between different types of states, such as different job categories. Yamaguchi (1991) and Petersen (1991) have provided accessible accounts of the methodology, emphasizing sociological applications, and Mayer and Tuma (1990) described a collection of case studies from social science. One important area of application of hazard rate models has been organizational birth and death processes; this is unique to sociology. Petersen (1995) extended the basic model further to multiple types of events where the events are interdependent—i.e. where the occurrence or nonoccurrence of one type of event affects the probability that the other type of event happens. An example is the relationship between becoming unemployed and getting

divorced. Xie (2000) has discussed the roots of event-history analysis in demography and life table analysis.

Uses of the Cox model in medicine have tended to treat the baseline hazard nonparametrically, but in social science it has sometimes been found useful to model it parametrically. For example, Yamaguchi (1992) analyzed permanent employment in Japan where the surviving fraction (those who never change jobs) and its determinants are of key interest; he found that covariates were associated both with the timing of job change and with the surviving fraction. Yamaguchi and Ferguson (1995) provide another application of this idea to the stopping and spacing of child births.

Social science event-history data are often recorded in discrete time—for example, by year—either because events tend to happen at particular times of year (e.g., graduation), or because of measurement constraints. As a result, discrete-time event-history models have been popular (Allison 1982, 1984; Xie 1994), and in some ways are easier to handle than their continuous-time analogs. Ways of dealing with multilevel event-history data, smoothly time-varying covariates, and other complications have been introduced in this context (e.g., Raftery, Lewis, and Aghajanian 1995; Fahrmeir and Knorr-Held 1997).

This basic framework has also been found useful to model a different kind of phenomenon: that of diffusion of innovations and social influence. Burt (1987) provided a theoretical framework for this work, and the extended event-history framework proposed for modeling it was developed by Marsden and Podolny (1990), Strang (1991), and Strang and Tuma (1993). A different approach, using accelerated failure-time models rather than proportional hazards models, was developed by Diekmann (1989) and Yamaguchi (1994).

One problem with social science event-history data is that dropping out can be related to the event of interest. For example, people may tend to leave a study shortly before a divorce, which will play havoc with estimation of divorce rates. The problem seems almost insoluble at first sight, but Hill (1997) produced an elegant solution using the Shared Unmeasured Risk Factor (SURF) model of Hill, Axinn, and Thornton (1993). The basic trick is to observe that, although one does not know which of the people who dropped out actually got divorced soon afterward, one can estimate which ones were most at risk of divorcing. One can then use this information to adjust the empirical divorce rates in the study by modeling divorce and dropout simultaneously.

3.4. *Binary Dependent Variables*

The term "limited dependent variable" is usually used to refer to a scalar dependent variable in a regression model, the set of whose possible values is restricted in a way that violates the assumptions of normal linear regression too severely for it to be used. The canonical example is binary dependent variables; others include nominal, ordinal, and compositional variables, and, in some contexts, variables that are constrained to be positive.

Limited dependent variables, especially binary variables, arise frequently in social research, and many articles in leading sociological journals use models and methods specifically developed for this situation. Nevertheless, much of the methodological development in this area has come from disciplines other than sociology. However, sociologists have played a major role in expositing, adapting, and synthesizing these methods; for example, see the books by Long (1997) and Xie and Powers (2000).

For binary responses, the method of choice in sociology in the past 20 years has been logistic regression. Much of the early development was for medical applications (Cornfield 1951, 1962; Truett et al. 1967), and the monograph by Cox (1970) helped to introduce the methods to a wide audience. The advent of generalized linear models (Nelder and Wedderburn 1972), and the recognition that logistic regression is a special case, as well as the development of the associated GLIM software (Baker and Nelder 1977), helped to make logistic regression a standard tool in many disciplines, particularly in the social and health sciences. Some version or descendant of the GLIM software is now included in most major commercial statistical packages.

Logistic regression is not the only possible model for regression with binary responses. Ordinary linear regression gives similar results if most of the probabilities are far enough from 0 and 1 (say between 0.1 and 0.9). Logistic regression is more "correct" than linear regression since, for example, it constrains fitted probabilities to lie between 0 and 1. Nevertheless, in the 1970s and 1980s there was a debate about whether logistic regression is really needed, given that it is more complex to estimate and needs more computer time than linear regression. The subsequent increase in computer speed made the additional computer time negligible, and the debate was settled in favor of logistic regression.

Another alternative is probit regression, in which the dependent variable is assumed to arise by truncating an unobserved normal random variable whose expectation depends linearly on the independent variables. This is soundly based and easy to estimate, because it is also a generalized linear model and so can be estimated using GLIM; it tends to give results that are very similar to those from logistic regression. However, in sociology, as in many other disciplines, it has lost out to logistic regression, perhaps because of the appealing interpretation of the logistic regression coefficients as odds ratios. There has been a revival of interest in probit regression recently among statisticians. This is because it is defined in terms of latent variables, and so can be included relatively easily as a component in more complex Bayesian models that are estimated using Markov chain Monte Carlo methods (e.g., Albert and Chib 1993).

A further alternative is complementary log-log regression, in which $\log(-\log(p))$ is assumed to be a linear combination of independent variables, where p is the conditional probability of the event of interest, given the independent variables. This is also a generalized linear model and so is easy to estimate. It can fit much better than logistic regression, and often gives quite different predicted probabilities, particularly for more extreme values of the independent variables. One example of this is the Irish educational transition data discussed by Raftery and Hout (1985); see Kass and Raftery (1995).

The introduction of two-sided logit models by Logan (1996, 1997) was an important development. This recognizes that in many situations in social life where individuals choose between different outcomes, there are two types of force in play: the preferences and attributes of the individual, and those of the possible choices. For example, in the labor market, which job an individual ends up in depends not only on his or her own attributes and preferences or utilities, but also on those of the other candidates in the job market, and those of the available employers and jobs. Logan's approach is to model both of these processes explicitly and simultaneously, and to explain the final labor market outcomes in terms of the interaction between them. The model can be estimated using either individual-level data or data aggregated into a cross-classification.

3.5. *Other Limited Dependent Variables*

Logistic regression has been extended to nominal dependent variables with more than two categories; for example, see Hosmer and Lemeshow

(1989). Maximum-likelihood estimation of the resulting multinomial logistic regression model is relatively straightforward, and software to do it is available. Begg and Gray (1984) have shown that this can be very well approximated by an appropriately set up (binary) logistic regression; see also Hosmer and Lemeshow (1989). Logistic regression has also been extended to ordinal dependent variables; for example, see McCullagh and Nelder (1989) and Agresti (1990).

Another important kind of limited dependent variable arises when the variable is positive, but has a nonnegligible probability of being exactly equal to zero. One example is income from work: Some people are out of the labor force or unemployed and have zero income from work, while all others have positive income. Data of this kind have often been analyzed using the Tobit model of Tobin (1958). In this model, it is assumed that those with zero income actually have an unobserved negative income, and the true income (now assumed to be capable of taking all positive and negative values) is modeled using ordinary linear regression.

The Tobit model in its original form seems rather unsatisfactory. For one thing, the postulated unobserved value does not exist: those who have zero income actually do have zero income (ignoring measurement error), not some unobserved negative income. Also, and perhaps more seriously, the model assumes that the mechanism determining whether or not someone has income from work is essentially the same as the one that determines how much he or she earns. It could easily be the case, however, that the mechanism that determines whether or not individuals are in the labor force is quite different from the mechanism that determines how much they earn if they are in the labor force, and with the Tobit model it is hard to make this distinction.

The Tobit model was developed before the widespread availability of specific methods for binary dependent variables. Now, however, there is a simple alternative approach that avoids the problems with the Tobit model. One simply models the data in two steps. In the first step, the dependent variable is whether or not the dependent variable is zero, and this is modeled using probit regression. Then in the second step, the dependent variable is the amount earned and only individuals with positive earnings are included. This is the standard sample selection model, which led to the development of the Heckman (1979) two-stage estimator. Amemiya (1985) calls this the Type II Tobit model. Winship and Mare (1992) review subsequent developments in this area.

A further kind of limited dependent variable arises in the analysis of compositional data. Here the dependent variable is a vector of positive values that sum up to one and consist of proportions. An example is the analysis of household budgets: the response is a vector, each element of which is the proportion of total household expenditure spent on some category, such as rent, food, utilities, education, and so on. One's first idea might be to model each proportion separately using regression, or perhaps to use a multivariate regression method that takes account of the correlation between the different responses. These methods do not work, however, because of the constraint that the responses add up to one, and so standard distributional assumptions do not apply. The observations lie on a *simplex*, the high-dimensional analog of the triangle, not on the full Euclidean space. A literature studying this situation has been summarized in part by Aitchison (1986); his main recommendation is to first transform the p-dimensional vector of proportions to a $(p-1)$-dimensional vector on the full Euclidean space using the multivariate logistic transform, and then to proceed using standard methods.

3.6. *Multilevel Models*

Multilevel models extend the regression models and their generalizations to situations where individual-level outcomes depend not just on individual-level covariates but also on social context. Much of the development in the social sciences has been in the context of education. A canonical example is where the individual-level outcomes are grades or test scores, and the contexts are the class, the school, the school district, the state, or some subset of these.

Often there is interest in the situation where the effect of an individual-level attribute, such as household income, depends on the context. For example, it might be hypothesized that in some schools the effects on test scores of inequalities due to differences in household income would be less than in other schools. The simplest approach to modeling such situations, with a view to estimating and testing the hypothesized effects, is via a fixed effects multilevel model. Suppose that y_i is the outcome for student i who attends school $s(i)$, where there are S schools represented in the data, and that x_i is his or her family income. Then a simple fixed effects model is

$$y_i = \alpha + \beta_{s(i)} x_i + \varepsilon_i, \tag{3}$$

where $\beta_{s(i)}$ is the effect of household income on test score in student i's school, and $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma_\varepsilon^2)$. There is a different regression coefficient β_j for each school j. This model can be estimated by ordinary least squares regression; for example, see Boyd and Iversen (1979) and Blalock (1984).

There are several difficulties with this model. One is that the number of parameters to be estimated, equal to $(S + 2)$, is large if there are many contexts (schools) involved, and so the model is hard both to estimate accurately and to interpret. Another is that, if the number of students from a particular school is small, and the estimated regression coefficient for that school is extreme relative to the estimates for the other schools, the resulting estimate is likely to be poor. This can be a problem, as it is often precisely these more extreme estimates that are of most interest.

There has been a great deal of work on overcoming these difficulties, and analogous ones in more complex and realistic multilevel situations, using random-effects models. In a simple formulation, (3) is supplemented by

$$\beta_j = \psi + \delta_j, \tag{4}$$

where $\delta_j \overset{\text{iid}}{\sim} N(0, \sigma_\delta^2)$. Combining (3) and (4) we get

$$y_i = \alpha + \psi x_i + u_i, \tag{5}$$

where

$$u_i \overset{\text{indep}}{\sim} N(0, \sigma_\varepsilon^2 + \sigma_\delta^2 x_i^2).$$

Equation (5) differs from (3) in having only four parameters to be estimated, instead of $(S + 2)$, and also in that the error variances differ, and depend on the value of the independent variable. One consequence is that the estimated "school effects" tend to be less extreme. It has been shown in several contexts that less extreme "shrunken" estimates such as those tend to be better on average (e.g., Morris 1983).

The basic idea of random-effects multilevel models goes back at least to Lindley and Smith (1972), who introduced the idea in a Bayesian context. Many different names have been used for the general class of models, including multilevel models, hierarchical models, random-effects models, variance component models, contextual models, random-coefficient models, and parametric empirical Bayes models. The area has benefited from a superb level of expository writing; for example, see Bock

(1989), Bryk and Raudenbush (1992), Longford (1993), DiPrete and Forristal (1994), Goldstein (1995), and Snijders and Bosker (1999). It has also spawned easy to use software, including HLM, MLn and VARCL, which has helped to spread the ideas.

Many of the applications have been in education, but there have been important applications in other areas of sociology. One successful application that helped to spread the methodology arose in demography, to modeling fertility decline (Mason et al. 1983; Entwisle et al. 1985, 1986, 1989; Wong and Mason 1985). Another fruitful area of application is meta-analysis—i.e., the pooling of results from different studies (Hedges and Olkin 1985; Goldstein et al. 2000).

The model can be estimated by maximum likelihood using the EM algorithm, viewing the random effects as "missing data" (Dempster, Laird, and Rubin 1977). The Bayesian formulation has proved useful in recent years, particularly for going beyond the hierarchical linear model, of which (5) is an example, to other more complex situations, such as multilevel models with limited dependent variables, event-history outcomes, multivariate outcomes, and so on. This has proved quite amenable to estimation using Markov chain Monte Carlo methods (e.g., Gelman et al. 1995; Daniels and Gatsonis 1999). Recent social science applications include Bradlow and Zaslavksy (1999), Boatwright et al. (1999), Datta et al. (1999), and Elliott and Little (2000). This seems to be a fruitful area for future research.

3.7. *Missing Data*

Missing data are pervasive in social science. By far the most common approach to dealing with the problem has been listwise deletion, in which cases with missing data on any of the relevant variables are removed from the analysis. Sometimes variables with a great deal of missing data are removed from the analysis as well. This works well as long as it does not lead to too many cases being removed: unbiased parameter estimators remain unbiased given that the missing data are missing at random, and the main problem is the loss of precision due to the reduction in the amount of data.

However, this approach starts to break down if the number of variables is considerable and the amount of missing data significant, as then much of the data can end up being removed. Various ways around this problem have been tried. One of these, mean imputation, in which the

missing value is replaced by the mean of the variable over the cases for which it is observed, can lead to biased estimates and is not to be recommended; unfortunately, it is frequently used, and it is even available in some widely distributed commercial software. Single imputation, also called regression imputation, consists of replacing the missing value by its conditional expectation given the values of the other variables for the case, estimated by regression. This gives unbiased estimates but tends to underestimate standard errors and other measures of uncertainty, to an extent that increases with the amount of missing information.

A consensus seems to be building that the method of choice for missing data is multiple imputation (Rubin 1977). This consists of simulating several replicates of the missing data from an approximate conditional or posterior distribution of the missing data given the observed data. These can then be combined to provide a composite inference that takes into account uncertainty about the missing data, does not discard any data, and is relatively easy to use (Little and Rubin 1987; Rubin 1987, 1996). This consensus is not a total one, and multiple imputation has been criticized and alternative suggestions made (e.g. Fay 1996; Rao 1996). It is possible, but more complicated, to specify a model for the missing data, and to compute maximum-likelihood estimates for the model (regression) parameters using the EM algorithm, taking account of the missing data and of uncertainty about them (Little and Rubin 1989; Little 1992).

The motivation for multiple imputation was Bayesian, and the resulting inferences are approximately Bayesian. Recently, a more exact Bayesian approach to this problem has been developed using Markov chain Monte Carlo (e.g. Schafer 1997). This extends multiple imputation by allowing one to simulate values of the missing data and of the parameters at the same time, to yield a sample from a posterior distribution of the parameters that takes full account of the missingness. This yields more accurate estimates and statements of uncertainty than Rubin's original version of multiple imputation, but it is also more cumbersome to implement.

Multiple imputation relies for its validity on the assumption that whether or not a particular value is missing is in some sense random and independent of the other data. The technical term *missing completely at random (MCAR)* was coined to denote the situation in which missingness is statistically independent of all the data, observed and unobserved. It turns out, fortunately, that this rather demanding assumption does not have to be met for multiple imputation to be valid. Instead, the missingness needs only to be conditionally independent of the unobserved data given

the observed data, a condition technically referred to as *missing at random (MAR)*. This latter condition holds, at least approximately, in many situations. It does not hold, however, if the missingness is related to the missing data themselves (e.g., if people with higher incomes are more likely to refuse to say what their income is). We have previously discussed one approach to this more difficult problem of "nonignorable missingness" in the specific case of event-history data, using the SURF model.

3.8. *Causality*

The goal of much of the regression and other statistical modeling that we have been discussing is, at least implicitly, to make statements about the mechanisms that underlie social life, social behavior, and social structure. In other words, to make causal statements. Statisticians, on the other hand, have tended to avoid the language of causality, cautioning that statistical models can show association between variables but cannot prove that the association is causal in origin.

The regression approach to causality has loomed large in social science because it seems to fit well with how empirical social researchers proceed. Much of (social) science proceeds by a researcher positing a causal theory of how and why a phenomenon occurs, implying that the presence of some attribute X causes an outcome Y. Data on observed values of X and Y are then collected. If a correlation between X and Y is observed, it provides some support for the causal theory but does not demonstrate it because there are other possible explanations of the correlation, notably: (a) Y might be causing X instead of the other way round, or (b) some third (set of) attribute(s) Z might be causing both X and Y.

The most common approach in these circumstances is to collect time-ordered or longitudinal data on X and Y to try to exclude (a), and to collect data on as many hypothesized common causes Z as possible to try to render (b) less plausible. This is done by "controlling" for Z—i.e., by assessing whether X and Y remain correlated when cases with each value of Z are considered separately. If Z can take many possible values (e.g., because it consists of several variables, or of variables that can take many values), this will not be feasible, and instead a regression model is built that represents the relationships in a more parsimonious way. If the "effect" of X on Y remains significant after controlling for Z, that is taken as evidence for the posited causal theory. It does not provide a conclusive dem-

onstration of X causing Y, however. For example, there might be other Z variables that we could not measure or did not think of.

When there is some additional causal information—such as the presence of an independent variable that is known to be causally related to one of X and Y but not to the other—causal inferences can sometimes be made. The basic approach is instrumental variables estimation, and this is a major topic in econometrics, but I will not discuss it further here.

Several scientists have been trying to make the case that one *can* infer causation from observational data in the absence of additional causal information, describing methods for doing so, and giving examples of its being done. This contention remains controversial. Two primary approaches to this task have been taken: the structural equation or graphical model approach and the counterfactual approach.

The first of these traditions of causal inference is that of structural equation modeling, or, more recently, graphical models. This tradition is motivated by the effort to infer causal structure from the multivariate (perhaps simply cross-sectional) structure of data. Perhaps the boldest claims about the possibility of doing this were made by Spirtes, Glymour, and Scheines (1993), drawing in part on work by Blalock (1961) and Costner (1969). They argued there that while the saying "correlation does not imply causation" is clearly true for two variables, it is not necessarily true for three or more variables. As the simplest example, they considered the case where the correlation structure of three variables, X, Y, and Z, is of the form $X - Y - Z$—i.e., X and Z are both correlated with Y but uncorrelated with each other. They pointed out that in this case, most people would agree that the causal structure of the data is of the form $X \rightarrow Y \leftarrow Z$, and they gave conditions under which this inference would be correct.

Extending this work, Spirtes et al. (1998) considered linear structural equation models, and asked several questions that arise. If there is a causal model that fits the data well, are there other equivalent models that imply the same covariance structure but a different path diagram, and if so, how many are there? Given that there are equivalent models, is it possible to extract the features common to all of them? When does a nonzero partial regression coefficient correspond to a nonzero coefficient in a structural equation? They provided answers to some of these questions using the key property of *d-separation*, defined by Verma and Pearl (1988). This can be viewed as a generalization of the concept of conditional independence. This makes it possible to read causal relations off the graph.

The second major current approach to causal inference is the counterfactual one. This starts from the idea that the randomized experiment with perfect compliance and no missing data is the gold standard for estimating the causal effects of treatment interventions. In social science, randomized experiments are sometimes done to estimate treatment effects—e.g., the effects of social programs. However, unlike randomized experiments in some other areas of science, such experiments suffer from the problem of noncompliance—some subjects refuse the treatment to which they are assigned. These experiments also tend to suffer from missing data.

The counterfactual approach to estimating causal effects from such experiments was first proposed by Rubin (1974) in the context of what later became known as the Rubin causal model. An accessible description of this framework was provided by Holland (1986); see also Manski (1993, 1995), Manski and Nagin (1998), and Heckman and Hotz (1989). This approach was illustrated by Barnard et al. (1998), who described methods for dealing with both noncompliance and missing data in this framework, illustrating their points with issues from the analysis of the Milwaukee Parental Choice Program, a natural randomized experiment.

Sobel (1990, 1994, 1995, 1997, 1998) has investigated the application of the counterfactual framework to observational data, which is more common in sociology than the (imperfect) randomized experiments that Rubin and his collaborators have considered. Sobel argues that when using data from observational studies, sociologists should attempt to identify causes, and then think about the covariates that would justify invoking the assumption of conditional random assignment, and attempt to measure these in the study. In Sobel (1998) he applied his reasoning to an attainment model of Featherman and Hauser (1976), concluding that the "effects" of family background on educational attainment and occupational achievement should not be viewed as causal. Sobel's conclusion in what would often be regarded as a rather clear-cut case of a causal effect suggests that few observational studies in sociology would meet his criteria for causal inference to be possible. Simplistically put, this is because one can rarely be sure that there is no unmeasured common cause out there. This is a useful caveat, as is his detailed description of the relatively rare cases when causal inference from observational studies will be possible.

However, much of sociology is about marshaling evidence for *competing* causal explanations, and observational studies *can* allow one to do this, regardless of whether or not they allow one to show any causal expla-

nation to be correct in an absolute sense. Such studies do provide a basis for saying which of the current causal theories is best supported by the data. The most common way of doing this is by testing one or several coefficients in a regression-type model for significance. This has the limitation that it can be used only to compare pairs of theories that correspond to nested statistical models. Often, however, competing theories do not neatly fit inside one another in this way, but instead correspond to quite different ways of explaining a phenomenon, and so do not correspond to nested hypotheses. In this case, standard statistical significance testing becomes difficult. However, Bayes factors can still be used to make these comparisons (Kass and Raftery 1995; Raftery 1995).

The search for causal explanation, although widely accepted as the basis for much social research, is not uncontroversial. For example, Abbott (1998) argued that the regression model of causality, although dominant in American sociology, is too narrow and needs to be expanded to include broader concepts of explanation and to reinstate the central role of description. He put forward the historical narrative-based approach as one way of achieving a more compelling and interesting account of social life. He mentioned nonstatistical simulation models and cluster analysis as potentially useful methods in this context, allowing one to describe relational and spatial aspects of social life, as well as temporal ones. This kind of "noncausal" or "postcausal" thinking is an important ingredient in the development of the third generation of methods, to which we now turn.

4. THE THIRD GENERATION: NEW DATA, NEW CHALLENGES, NEW METHODS

4.1. *Social Networks and Spatial Data*

Social networks consist of sets of pairwise connections, such as friendships between adolescents (Udry and Bearman 1998), sexual relationships between adults (Laumann et al. 1994), or patterns of marriage exchange and political alliance across social groups (White 1963; Bearman 1997; Padgett and Ansell 1993). The analysis of data about such networks has a long history (Wasserman and Faust 1994). Frank and Strauss (1986) developed formal statistical models for such networks related to the Markov random field models used in Bayesian image analysis, and derived using the Hammersley-Clifford theorem (Besag 1974). This has led to the promising "p^*" class of models for social networks (Wasser-

man and Pattison 1996). An alternative approach to formal statistical modeling of social networks based on Goodman-type association models is due to Yamaguchi (1990).

Methods for the analysis of social networks have focused mostly on small data sets with complete data. In practical applications, however, such as the effect of sexual network patterns on the spread of sexually transmitted diseases (Morris 1997), the data tend to be large and very incomplete, and current methods are somewhat at a loss. This is the stage that pedigree analysis in statistical genetics was at some years ago, but the use of likelihood and MCMC methods have led to major progress since then (Thompson 1998). Social networks are more complex than pedigrees in one way, because pedigrees tend to have a tree structure, while social networks often have cycles, but progress does seem possible.

Most social data are spatial, but this fact has been largely ignored in sociological research. A major exception is Massey and Denton's (1993) study of residential segregation by race, reviving a much older sociological tradition of spatial analysis in American society (e.g., Duncan and Duncan 1957). More recently, the field of research on fertility and contraception in Asia (several major projects focused on China, Thailand, and Nepal) has been making fruitful use of satellite image and Geographic Information System (GIS) data (e.g., Entwisle et al. 1997).

More extensive use of spatial statistics in sociology seems likely. Spatial statistics has been making great progress in the past two decades. The two most fruitful approaches to modeling spatial dependence have turned out to be those based on geostatistics (Matheron 1971; Chilès and Delfiner 1999), and on Markov random fields (Besag 1974; Besag, York, and Mollié 1991). Geostatistics models spatial correlation taking account of distance explicitly. Markov random fields, on the other hand, are based on a notion of neighborhood: an observation is taken to depend directly on its neighbors, and to be conditionally independent of all other cases given its neighbors. Markov random fields seem promising for social data if they are fairly regularly spaced, but for unevenly spaced spatial units, geostatistics may find it easier to account for the spatial dependence. For social data, geographic distance may not be the most relevant; distances defined on the basis, for example, of flows of people or of information may be more germane for some applications. I do not know of any work on spatial statistical models based on distances of this type, however.

4.2. *Textual and Qualitative Data*

In its rawest form, a great deal of sociological data is textual—for example, interviews, answers to open-ended questions in surveys, and ethnographic accounts. How to analyze such data formally and draw inference from them remains a largely open question. Efforts at formal analysis have focused on standard content analysis, consisting mainly of counting words in the text in different ways. It seems likely that using the context in which words and clauses appear would yield better results. Promising recent efforts to do just this include Carley's (1993) map analysis, Franzosi's (1994) set theoretic approach, and Roberts's (1997) generic semantic grammar, but the surface has only been scratched. The human mind is very good at analyzing individual texts, but computers are not, at least as yet; in this way the analysis of textual data may be like other problems such as image analysis and speech recognition. A similar challenge is faced on a massive scale by information retrieval for the Web (Jones and Willett 1997), where most search engines are based on simple content analysis methods. The more contextual methods being developed in sociology might be useful in this area also.

Singer et al. (1998) have made an intriguing use of textual data analysis, blending quantitative and qualitative approaches. They took a standard unit-level data set with more than 250 variables per person, and converted them into written "biographies." They then examined the biographies for common features, and thinned them to more generic descriptions.

Another approach to the systematic analysis of some kinds of qualitative data has recently been pioneered by Raudenbush and Sampson (1999) under the name "ecometrics." Their work was motivated by the study of neighborhood characteristics that could be linked to crime, such as physical decay (e.g., abandoned buildings), physical disorder (e.g., graffiti), and social disorder (e.g., drug dealing on the street). A standard quantitative approach to this kind of problem has been to estimate neighborhood effects using aggregates of respondents from the neighborhood, but Raudenbush and Sampson argue persuasively that this does not provide independent or "objective" assessments of the environment based on direct observation. Their data consisted of videotapes and observer logs for about 23,000 street block segments (Sampson and Raudenbush 1999). They coded these data and developed a hierarchical model for assessing reliabilities and calculating physical and social disorder scales.

Raudenbush and Sampson place their work firmly in Reiss's (1971) framework of systematic social observation, defined to include explicit rules that permit replication, and means of observation that are independent of what is being observed. This seems important for formal analysis of and inference from qualitative, textual, and ethnographic data; the work of Carley (1993), Franzosi (1994) and Roberts (1997) is in this spirit.

Interestingly, Raudenbush and Sampson point out that the search for individual and ecological effects may overemphasize the individual component simply because the well-studied individual psychometric measures are likely to be better than the much less studied ecological ones. Indeed, I have noticed that in many sociological studies, the reported contextual and neighborhood effects are weak, and the point that this may be due to poor measures rather than to weak effects is interesting. Data of this kind cry out for spatial statistical analysis. Raudenbush and Sampson acknowledge this and list it as a topic for future research; their work to date has not accounted for spatial dependence.

4.3. *Narrative and Sequence Analysis*

Life histories are typically analyzed by reducing them to variables and doing regression and multivariate analysis, or by event-history analysis. Abbott and Hrycak (1990) argued that these standard approaches obscure vital aspects of a life history (such as a professional career) that emerge when it is considered as a whole. They proposed viewing life histories of this kind as analogous to DNA or protein sequences, using optimal alignment methods adapted from molecular biology (Sankoff and Kruskal 1983), followed up by cluster analysis, to detect patterns common to groups of careers. Stovel, Savage, and Bearman (1996) used these methods to describe changes in career systems at Lloyds Banks over the past century.

Subsequently, Dijkstra and Taris (1995) extended the ideas to include independent variables, and Abbott and Barman (1997) applied the Gibbs sampling sequence detection method of Lawrence et al. (1993), originally also developed for microbiology; this seems to work very well.

The approach is interesting, and there are many open statistical questions. These include questions about the alignment methodology—for example, how should the insertion, deletion, and replacement costs be determined? They also include questions about the clustering method: How many clusters are there? Which clustering method should be used? How should one deal with outliers? Perhaps a more explicitly model-based

approach would help to answer these questions. Cluster analysis was long a somewhat ad hoc collection of methods, and reformulating it so that it is based on formal statistical models has helped provide principled answers to some of these questions in other contexts (e.g., Banfield and Raftery 1993; Fraley and Raftery 1998). An alternative approach to analyzing sequence data based on log-linear models has been developed by Yamaguchi and Kandel (1998).

4.4. *Simulation Models*

Another way to represent a social process in more detail is via a macro- or microsimulation model. Such models are often deterministic and quite complicated, representing systems by different compartments that interact, and each compartment by a set of differential or difference equations. They have been used, for example, to explore the implications of different theories about how domestic politics and war interact (Hanneman, Collins, and Mordt 1995), the social dynamics of collective action (Kim and Bearman 1997), and the role of sexual networks in the spread of HIV (Morris 1997 and references therein).

A difficulty with such models is that ways of estimating the many parameters involved, of assessing the fit of the model, and of comparing competing models are not well established; all this tends to be done by informal trial and error. Methods being developed to put inference for such models on a solid statistical footing in other disciplines may prove helpful in sociology as well (Guttorp and Walden 1987; Raftery, Givens, and Zeh 1995; Poole and Raftery 2000).

4.5. *Macrosociology*

Macrosociology deals with large entities, such as states and their interactions. As a result, the number of cases tends to be small, and the use of standard statistical methods such as regression is difficult. This was pointed out trenchantly by Ragin (1987) in an influential book. His own proposed alternative, qualitative comparative analysis, seems unsatisfactory because it does not allow for variability of any kind, and so is sensitive to small changes in the data and in the way the method is applied (Lieberson 1994).

One solution to the problem is to obtain an at least moderately large sample size, as Bollen and Appold (1993) were able to do, for example. Often, however, this is not possible, so this is not a general solution.

Another approach is to use standard regression-type models, but to do Bayesian estimation with strong prior information if available, which it often is from the practice, common in this area, of analyzing specific cases in great detail (Western and Jackman 1994). Bayes factors may also help, as they tend to be less stringent than standard significance tests in small samples and allow a calibrated assessment of evidence rather than forcing the rejection or acceptance of a hypothesis (Kass and Raftery 1995). They also provide a way of accounting for model uncertainty, which can be quite large in this context (Western 1996).

5. DISCUSSION

Statistical methodology has had a successful half-century in sociology, leading the way in providing models for cross-classifications, and developing well-adapted methods for unit-level data sets. This has contributed to the greatly improved level of scientific rigor in sociology today. New kinds of data and new challenges abound, and the area is ripe for statistical research.

What are the future directions? As is implicit in my categorization of generations, I feel that the questions posed by the types of data that have motivated the third generation of methods may well spark some of the most exciting developments in sociological methodology in the medium term. But there are others, particularly related to the kind of data that may emerge from current technological developments. For example, surveys carried out by giving computers to respondents and inviting them to respond online, perhaps sporadically or repeatedly over an extended period, may generate useful data with new methodological issues of repeated measures at unequal time intervals and missing data (or they may not work at all). More generally, the Web is generating vast amounts of social science data of new types, and developing methods for drawing valid conclusions from such data is bound to be a major future source of challenges.

One direction I would both predict and advocate is that future developments will be interdisciplinary, spanning the social sciences and beyond. This has not been the case for most of the twentieth century, during which one social science discipline after another made the leap to greater quantitative sophistication, but often in relative isolation from one another and from statistics as a whole. Psychology may have been the first to make this transition, with the work of Spearman and Thurstone early in the century, followed by economics, with the development of econometrics in

the 1930s and 1940s by Haavelmo, Tinbergen, the Cowles Commission, and others. Then sociology made its move in the 1960s, with the work of Blalock, Duncan, Goodman, and others that we have been discussing here. In the 1990s, it has been the turn of political science, led by Gary King, Larry Bartels, and others, who have been adopting and adapting modern statistical methodology to their discipline and developing new methods in the process.

The pattern in each of these disciplines has been similar. The quantitative transition has tended to focus on, and in some cases create, the most advanced statistical methods available at the time, and to spawn a dynamic cadre of methodologists, which in the case of the disciplines that made the transition the longest time ago, psychology and economics, have coalesced into their own quasi-disciplines of psychometrics and econometrics. Subsequent quantitative methodological development has been slower in each discipline, however, and has tended to remain tied to the methods that were at the cutting edge at the time of the quantitative transition. Sociology has not escaped this pattern: there quantitative work remains dominated by the methods first developed in the 1960s and early 1970s (structural equation models with latent variables, generalized linear models, event-history analysis via the Cox model), and has focused on developments and refinements of these methods. As I have discussed, there are good reasons for this, and it has had a very positive effect on the field as a whole. However, the statistical methods of the 1990s, particularly Bayesian analysis via Markov chain Monte Carlo, have been eagerly adopted by the cohorts of young political scientists going through the excitement and turmoil of their own quantitative revolution, but have been slower to penetrate sociology.

Now, in an academic world more interdisciplinary than that of previous decades, the opportunity is there for all the social science disciplines to break out of the disciplinary straitjacket and to move their quantitative methodologies forward together. Several major institutions have launched interdisciplinary centers and initiatives focused on quantitative social science methodology in the past few years, providing resources for doing just this. The University of Washington has just established a new Center for Statistics and the Social Sciences. Harvard's new Center for Basic Research in the Social Sciences emphasizes social statistics. The new Center for Spatially Integrated Social Science at the University of California–Santa Barbara is another example, with a focus on spatial statistics. UCLA's young Statistics Department grew out of social statistics,

and retains active interdisciplinary links to several social sciences. Columbia's new master's program in Quantitative Social Science is another interdisciplinary enterprise spanning the social sciences and statistics. At the University of Michigan, the new Quantitative Methodology Program is creating and reviving joint graduate programs between the Department of Statistics and several social science departments. These all join what is perhaps the most successful effort of this kind to date: the Social Statistics Department at the University of Southampton.

REFERENCES

Abbott, Andrew. 1998. "The Causal Devolution." *Sociological Methods and Research* 27:148–81.

Abbott, Andrew, and Emily Barman. 1997. "Sequence Comparison Via Alignment and Gibbs Sampling: A Formal Analysis of the Emergence of the Modern Sociological Article." *Sociological Methodology* 27:47–88.

Abbott, Andrew, and Alexandra Hrycak. 1990. "Measuring Sequence Resemblance," *American Journal of Sociology* 96:144–85.

Agresti, Alan. 1990. *Categorical Data Analysis.* New York: Wiley.

Aitchison, John. 1986. *The Analysis of Compositional Data.* London: Chapman and Hall.

Albert, James, and Siddartha Chib. 1993. "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association* 88:669–79.

Allison, Paul. 1982. "Discrete-Time Methods for the Analysis of Event Histories." *Sociological Methodology* 13:61–98.

———. 1984. *Event History Analysis.* Beverly Hills, CA: Sage.

Alwin, Duane F., and Robert M. Hauser. 1975. "The Decomposition of Effects in Path Analysis." *American Sociological Review* 40:37–47.

Amemiya, Takashi 1985. *Advanced Econometrics.* Cambridge, MA: Harvard University Press.

Arminger, Gerhard. 1998. "A Bayesian Approach to Nonlinear Latent Variable Models Using the Gibbs Sampler and the Metropolis-Hastings Algorithm." *Psychometrika* 63:271–300.

Baker, R. J., and John A. Nelder. 1977. *The GLIM System, Release 3, Generalized Linear Interactive Modeling.* Oxford, England: Numerical Algorithms Group.

Banfield, Jeffrey D., and Adrian E. Raftery. 1993. "Model-Based Gaussian and Non-Guassian Clustering." *Biometrics* 49:803–21.

Barnard, John, J. T. Du, Jennifer L. Hill, and Donald B. Rubin. 1998. "A Broader Template for Analyzing Broken Randomized Experiments." *Sociological Methods and Research* 27:285–317.

Bearman, Peter S. 1997. "Generalized Exchange." *American Journal of Sociology* 102:1383–415.

Beck, Nathaniel. 2000. "Political Methodology: A Welcoming Discipline." *Journal of the American Statistical Association* 95:651–54.

Becker, Mark P. 1989. "Using Association Models to Analyze Agreement Data: Two Examples." *Statistics in Medicine* 8:1199–207.

———. 1994. "Analysis of Cross-Classifications of Counts Using Models for Marginal Distributions: An Application to Trends in Attitudes on Legalized Abortion." *Sociological Methodology* 24:229–65.

Becker, Mark P., and I. Yang. 1998. "Latent Class Marginal Models for Cross-Classifications of Counts." *Sociological Methodology* 28:293–326.

Begg, Colin B., and R. Gray. 1984. "Calculation of Polytomous Logistic Regression Parameters Using Individualized Regressions." *Biometrika* 71:11–18.

Benzécri, J.-P. 1976. *L'Analyse des Données.* 2d ed. Paris: Dunod.

Bernhardt, Annette D., Martina Morris, and Mark S. Handcock. 1995. "Women's Gains or Men's Losses? A Closer Look at the Shrinking Gender Gap in Earnings." *American Journal of Sociology* 101:302–28.

Besag, Julian E. 1974. "Spatial Interaction and the Statistical Analysis of Lattice Systems" (with discussion). *Journal of the Royal Statistical Society*, Ser. B, 36:192–236.

Besag, Julian E., Jeremy York, and Annie Mollié. 1991. "Bayesian Image Restoration, with Two Applications in Spatial Statistics" (with discussion). *Annals of the Institute of Statistical Mathematics* 43:1–59.

Biblarz, Timothy J., and Adrian E. Raftery. 1993. "The Effects of Family Disruption on Social Mobility." *American Sociological Review* 58:97–109.

Biblarz, Timothy J., Adrian E. Raftery, and Alexander Bucur. 1997. "Family Structure and Social Mobility." *Social Forces* 75:1319–39.

Birch, M. W. 1963. "Maximum Likelihood in Three-Way Tables," *Journal of the Royal Statistical Society,* Ser. B, 25:220–33.

———. 1965. "The Detection of Partial Association, II: The General Case," *Journal of the Royal Statistical Society,* Ser. B, 27, 111–124.

Blalock, Hubert M. 1961. *Causal Inferences in Nonexperimental Research.* New York: W.W. Norton.

———. 1984. "Contextual-Effects Models: Theoretical and Methodological Issues." *Annual Review of Sociology* 10:353–72.

Blau, Peter M., and Otis Dudley Duncan. 1967. *American Occupational Structure.* New York: Free Press.

Boatwright, Peter, Robert McCullouch, and Peter Rossi. 1999. "Account-level Modeling for Trade Promotion: An Application of a Constrained Parameter Hierarchical Model." *Journal of the American Statistical Association* 94:1063–73.

Bock, R. D. 1989. *Multilevel Analysis of Educational Data.* San Diego, CA: Academic Press.

Bollen, Kenneth A. 1989. *Structural Equation Models with Latent Variables.* New York: Wiley.

Bollen, Kenneth A., and S. J. Appold. 1993. "National Industrial-Structure and the Global System." *American Sociological Review* 58:283–301.

Boyd, L. H., and G. R. Iversen. 1979. *Contextual Analysis: Concepts and Statistical Techniques.* Belmont, CA: Wadsworth.

Bradlow, Eric T., and Alan Zaslavsky. 1999. "A Hierarchical Latent Variable Model for Ordinal Data from a Customer Satisfaction Survey with 'No Answer' Responses." *Journal of the American Statistical Association* 94:43–52.

Browne, Michael W. 2000. "Psychometrics." *Journal of the American Statistical Association* 95:661–65.

Bryk, Anthony S., and Stephen W. Raudenbush. 1992. *Hierarchical Linear Models: Applications and Data Analysis Methods.* Newbury Park, CA: Sage.

Burt, Ronald S. 1987. "Social Contagion and Innovation: Cohesion Versus Structural Equivalence." *American Journal of Sociology* 92:1287–335.

Camic, Charles, and Yu Xie. 1994. "The Statistical Turn in American Social Science—Columbia University, 1890 to 1915." *American Sociological Review* 59:773–805.

Carley, Kathleen M. 1993. "Coding Choices for Textual Analysis: A Comparison of Content Analysis and Map Analysis." *Sociological Methodology* 23:75–126.

Celeux, Gilles, and Gérard Govaert. 1991. "Clustering Criteria for Discrete Data and Latent Class Models." *Journal of Classification* 8:157–76.

Charles, Maria, and David Grusky. 1995. "Models for Describing the Underlying Structure of Sex Segregation." *American Journal of Sociology* 100:931–71.

Chickering, D. Maxwell, and David Heckerman. 1997. "Efficient Approximations for the Marginal Likelihood of Bayesian Networks with Hidden Variables." *Machine Learning* 29:181–212.

Chilès, Jean-Paul, and Pierre Delfiner. 1999. *Geostatistics: Modeling Spatial Uncertainty.* New York: Wiley.

Clogg, Clifford C. 1992. "The Impact of Sociological Methodology on Statistical Methodology" (with discussion). *Statistical Science* 7:183–207.

———. 1995. "Latent Class Models." Pp. 311–59 in *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, edited by G. Arminger, C. C. Clogg and M. E. Soble. New York: Plenum.

Clogg, Clifford C., Scott R. Eliason, and R. J. Wahl. 1990. "Labor Market Experiences and Labor Force Outcomes." *American Journal of Sociology* 95:1536–76.

Cornfield, Jerome 1951. "A Method of Estimating Comparative Rates from Clinical Data; Application to Cancer of the Lung, Breast and Cervix." *Journal of the National Cancer Institute* 11:1269–75.

———. 1962. "Joint Dependence of the Risk of Coronary Heart Disease on Serum Cholesterol and Systolic Blood Pressure: A Discriminant Function Analysis." *Federation Proceedings* 21:58–61.

Costner, Herbert L. 1969. "Theory, Deduction and Rules of Correspondence." *American Journal of Sociology* 75:245–63.

Cox, David R. 1970. *The Analysis of Binary Data.* London: Chapman and Hall.

———. 1972. "Regression Models and Life Tables" (with discussion). *Journal of the Royal Statistical Society*, Ser. B, 34, 187–220.

Daniels, Michael J., and Constantine Gatsonis. 1999. "Hierarchical Generalized Linear Models in the Analysis of Variations in Health Care Utilization." *Journal of the American Statistical Association* 94:29–42.

Datta, G. S., Partha Lahiri, T. Maiti, and K. L. Lu 1999. "Hierarchical Bayes Estimation of Unemployment Rates for the States of the U.S." *Journal of the American Statistical Association* 94:1074–82.

De Graaf, Nan. 1991. "Distinction by Consumption in Czechoslovakia, Hungary and the Netherlands." *European Sociological Review* 7:267–90.

Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm" (with discussion). *Journal of the Royal Statistical Society*, Ser. B, 39:1–38.

Diekmann, Andreas. 1989. "Diffusion and Survival Models for the Process of Entry into Marriage." *American Journal of Mathematical Sociology* 14:31–44.

Dijkstra, W., and T. Taris. 1995. "Measuring the Agreement Between Sequences." *Sociological Methods and Research* 24:214–31.

DiPrete, Thomas A., and Jerry D. Forristal. 1994. "Multilevel Models: Methods and Substance." *Annual Review of Sociology* 20:331–57.

Duncan, Otis Dudley. 1961. "A Socioeconomic Index for All Occupations." Pp. 109–38 in *Occupations and Social Status*, edited by A. J. Reiss. New York: Free Press.

———. 1966. "Path Analysis." *American Journal of Sociology* 72:1–16.

———. 1975. *An Introduction to Structural Equation Models*. New York: Academic Press.

———. 1979. "How Destination Depends on Origin in the Occupational Mobility Table." *American Journal of Sociology* 84:793–803.

Duncan, Otis Dudley, and Beverly Duncan. 1957. *The Negro Population of Chicago*. Chicago: University of Chicago Press.

Durkheim, Emile. 1897. *Le Suicide*. Paris: Alcan. Translated by G. Simpson and J. A. Spaulding. 1951. New York: Free Press.

Eisenberg, Theodore. 2000. "Empirical Methods and the Law." *Journal of the American Statistical Association* 95:665–70.

Elliott, Michael R., and Roderick J. A. Little. 2000. "A Bayesian Approach to Combining Information from a Census, a Coverage Measurement Survey, and Demographic Analysis." *Journal of the American Statistical Association* 95:351–62.

Entwisle, Barbara, John B. Casterline, and H. A. A. Sayed. 1989. "Villages as Contexts for Contraceptive Behavior." *American Sociological Review* 54:1019–34.

Entwisle, Barbara, and William M. Mason. 1985. "Multilevel Effects of Socioeconomic Development and Family Planning Programs on Children Ever Born." *American Journal of Sociology* 91:616–49.

Entwisle, Barbara, William M. Mason, and A. I. Hermalin. 1986. "The Multilevel Dependence of Contraceptive Use on Socioeconomic Development and Family Planning Program Strength." *Demography* 23:199–215.

Entwisle, Barbara, Ronald R. Rindfuss, S. J. Walsh, T. P. Evans, and Sara R. Curran. 1987. "Geographic Information Systems, Spatial Network Analysis, and Contraceptive Choice." *Demography* 34:171–87.

Fahrmeir, Ludwig, and Leo Knorr-Held. 1997. "Dynamic Discrete-Time Duration Models: Estimation via Markov Chain Monte Carlo." *Sociological Methodology* 27:417–52.

Fay, Robert E. 1996. "Alternative Paradigms for the Analysis of Imputed Survey Data." *Journal of the American Statistical Association* 91:490–98.

Featherman, David L., and Robert M. Hauser. 1976. "Sexual Inequalities and Socioeconomic Achievement in the U.S., 1962–1973." *American Sociological Review* 41:462–83.

Featherman, David L., and Gillian Stevens. 1982. "A Revised Socioeconomic Index of Occupational Status: Applications in Analysis of Sex Differences in Attainment." Pp. 93–129 in *Measures of Socioeconomic Status*, edited by M. Powers. Boulder, CO: Westview.

Fienberg, Stephen E. 2000. "Contingency Tables and Log-Linear Models: Basic Results and New Developments." *Journal of the American Statistical Association* 95:643–47.

Fienberg, Stephen E., and William M. Mason. 1979. "Identification and Estimation of Age-Period-Cohort Effects in the Analysis of Discrete Archival Data." *Sociological Methodology* 10:1–67.

Fraley, Christina, and Adrian E. Raftery. 1998. "How Many Clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis." *Computer Journal* 41:578–88.

Frank, Ove, and David Stauss. 1986. "Markov Graphs." *Journal of the American Statistical Association* 81:832–42.

Franzosi, Roberto. 1994. "From Words to Numbers: A Set Theory Framework for the Collection, Organization and Analysis of Narrative Data." *Sociological Methodology* 24:105–36.

Freeman, John, Glenn Carroll, and Michael T. Hannan. 1983. "The Liability of Newness: Age Dependence in Organizational Death Rates." *American Sociological Review* 48:692–710.

Freedman, David A. 1987. "As Others See Us" (with discussion). *Journal of Educational Statistics*, 12, 101–223.

Ganzeboom, Harry B. G., Ruud Luijkx, and Donald J. Treiman. 1989. "Intergenerational Class Mobility in Comparative Perspective." *Research in Social Stratification and Mobility* 9:3–79.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 1995. *Bayesian Data Analysis*. New York: Chapman and Hall.

Gilks, Walter R., Sylvia Richardson, and David J. Spiegelhalter, eds. 1996. *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Glass, David V. 1954. *Social Mobility in Britain*. Glencoe, IL: Free Press.

Goldstein, Harvey. 1995. *Multilevel Models in Educational and Social Research*, 2d ed. London: Griffin.

Goldstein, Harvey, M. Yang, R. Omar, R. Turner, and S. Thompson. 2000. "Meta-Analysis Using Multilevel Models with an Application to the Study of Class Size Effects." *Applied Statistics* 49:399–412.

Goodman, Leo A. 1971. "The Analysis of Multidimensional Contingency Tables: Stepwise Procedures and Direct Estimation Methods for Building Models for Multiple Classifications." *Technometrics* 13:33–61.

———. 1974a. "The Analysis of Systems of Qualitative Variables When Some of the Variables Are Unobservable." *American Journal of Sociology* 79:1179–259.

———. 1974b. "Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models." *Biometrika* 61:215–31.

———. 1979. "Simple Models for the Analysis of Association in Cross-Classifications Having Ordered Categories." *Journal of the American Statistical Association* 74:537–52.

———. 1985. "The Analysis of Cross-Classified Data Having Ordered and/or Unordered Categories." *Annals of Statistics* 13:10–69.

Goodman, Leo A., and Michael Hout. 1998. "Statistical Methods and Graphical Displays for Analyzing How the Association Between Two Qualitative Variables Differs Among Countries, Among Groups, or Over Time: A Modified Regression-Type Approach" (with discussion). Pp. 175–262 in *Sociological Methodology 1998*, edited by Adrian E. Raftery. Cambridge, MA: Blackwell Publishers.

Grusky, David B., and Robert M. Hauser. 1984. "Comparative Social Mobility Revisited: Models of Convergence and Divergence in Sixteen Countries." *American Sociological Review* 49:19–38.

Guttorp, Peter, and Andrew T. Walden. 1987. "On the Evaluation of Geophysical Models." *Geophysical Journal of the Royal Astronomical Society* 91:201–10.

Haberman, Shelby J. 1979. *Analysis of Qualitative Data*, vol. 2. New York: Academic Press.

Hagenaars, Jacques A. 1988. "Latent Structure Models with Direct Effects Between Indicators: Local Dependence Models." *Sociological Methods and Research* 16: 379–406.

———. 1990. *Categorical Longitudinal Data: Log-Linear Panel, Trend and Cohort Analysis.* Newbury Park, CA: Sage.

Handcock, Mark S., and Martina Morris. 1998. "Relative Distribution Methods." *Sociological Methodology* 28, 53–98.

———. 1999. *Relative Distribution Methods in the Social Sciences.* New York: Springer-Verlag.

Hanneman, R. A., Randall Collins, and Gabriele Mordt. 1995. "Discovering Theory Dynamics by Computer Simulation: Experiments on State Legitimacy and Imperialist Capitalism." *Sociological Methodology* 25:1–46.

Hauser, Robert M., and John R. Warren. 1997. "Socioeconomic Indexes for Occupations: A Review, Update and Critique." *Sociological Methodology* 27:177–298.

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrika* 47:153–61.

Heckman, James J., and V. Joseph Hotz. 1989. "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training" (with discussion). *Journal of the American Statistical Association* 84:862–80.

Hedges, Larry J., and Ingram Olkin. 1985. *Statistical Methods for Meta-Analysis.* New York: Academic Press.

Hill, D. H. 1997. "Adjusting for Attrition in Event-History Analysis." *Sociological Methodology* 27:393–416.

Hill, D. H., W. G. Axinn, and A. Thornton. 1993. "Competing Hazards with Shared Unmeasured Risk Factors." *Sociological Methodology* 23:245–77.

Holland, Paul. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:945–70.

Hosmer, David W., and Stanley Lemeshow. 1989. *Applied Logistic Regression.* New York: Wiley.

Hout, Michael. 1983. *Mobility Tables.* Beverly Hills, CA: Sage.

———. 1984. "Status, Autonomy and Training in Occupational Mobility." *American Journal of Sociology* 89:1379–409.

———. 1988. "Expanding Universalism, Less Structural Mobility: The American Occupational Structural in the 1980s." *American Journal of Sociology* 93:1358–400.

Jones, K. S., and Willett, P. 1997. *Readings in Information Retrieval.* San Francisco: Morgan Kaufman.

Jöreskog, Karl G. 1973. "A General Method for Estimating a Linear Structural Equation System." Pp. 85–112 in *Structural Equation Models in the Social Sciences*, edited by A. S. Goldberger and O. D. Duncan. New York: Seminar.

Kalmijn, M. 1991. "Status Homogamy in the United States." *American Journal of Sociology* 97:496–523.

Kass, Robert E., and Adrian E. Raftery. 1995. "Bayes Factors." *Journal of the American Statistical Association* 90:773–95.

Kass, Robert E., and Larry Wasserman. 1995. "A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion." *Journal of the American Statistical Association* 90:928–34.

Kim, Hyojoung, and Peter S. Bearman. 1997. "The Structure and Dynamics of Movement Participation." *American Sociological Review* 62:70–93.

Koster, Jan. 1996. "Markov Properties of Non-Recursive Causal Models." *Annals of Statistics* 24:2148–77.

Kuo, H. H. D., and Robert M. Hauser. 1996. "Gender, Family Configuration, and the Effect of Family Background on Educational Attainment." *Social Biology* 43:98–131.

Lang, Joseph B., and Alan Agresti. 1994. "Simultaneously Modeling Joint and Marginal Distributions of Multivariate Categorical Responses." *Journal of the American Statistical Association* 89:625–32.

Laumann, Edward O., J. Gagnon, R. Michael, and S. Michaels. 1994. *The Social Organization of Sexuality*. Chicago: University of Chicago Press.

Lawrence, C. E., S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wooton. 1993. "Detecting Subtle Sequence Signals." *Science* 262:208–14.

Lazarsfeld, Paul F. 1950. "The Logical and Mathematical Foundation of Latent Structure Analysis." Pp. 362–412 in *Studies in Social Psychology in World War II.* Vol. 4, *Measurement and Prediction*, edited by E. A. Schulman, P. F. Lazarsfeld, S. A. Starr, and J. A. Clausen. Princeton, NJ: Princeton University Press.

Lazarsfeld, Paul F., and Neil W. Henry. 1968. *Latent Structure Analysis.* Boston: Houghton Mifflin.

Lieberson, Stanley L. 1994. "More on the Uneasy Case for Using Mill-Type Methods in Small-N Comparative Studies." *Social Forces* 72:1225–37.

Lindley, Dennis V., and Adrian F. M. Smith. 1972. "Bayes Estimates for the Linear Model" (with discussion). *Journal of the Royal Statistical Society*, Ser. B, Methodological, 34:1–41.

Little, Roderick J. A. 1992. "Regression with Missing *X*'s: A Review." *Journal of the American Statistical Association* 87:1227–37.

Little, Roderick J. A., and Donald B. Rubin. 1987. *Statistical Analysis with Missing Data.* New York: Wiley.

———. 1989. "The Analysis of Social Science Data with Missing Values." *Sociological Methods and Research* 18:292–326.

Logan, John A. 1996. "Opportunity and Choice in Socially Structured Labor Markets." *American Journal of Sociology* 101:114–60.

———. 1997. "Estimating Two-Sided Logit Models." *Sociological Methodology* 28:139–73.

Long, J. Scott. 1984a. *Confirmatory Factor Analysis.* Newbury Park, CA: Sage.

———. 1984b. *Covariance Structure Models.* Newbury Park, CA: Sage.

———. 1997. *Regression Models for Categorical and Limited Dependent Variables.* Thousand Oaks, CA: Sage.

Longford, Nicholas. 1993. *Logistic Regression with Random Coefficients.* Princeton, NJ: Educational Testing Service.

Manski, Charles F. 1993. "Identification Problems in the Social Sciences." Pp. 1–56 in *Sociological Methodology 1993*, edited by Peter V. Marsden. Oxford, England: Blackwell Publishers.

———. 1995. *Identification Problems in the Social Sciences.* Cambridge, MA: Harvard University Press.

Manski, Charles C., and Daniel S. Nagin. 1998. "Bounding Disagreements About Treatment Effects: A Case Study of Sentencing and Recidivism." Pp. 99–137 in *Sociological Methodology 1998*, edited by Adrian E. Raftery. Cambridge, MA: Blackwell Publishers.

Marsden, Peter V., and Joel Podolny. 1990. "Dynamic Analysis of Network Diffusion Processes." Pp. 197–214 in *Social Networks Through Time*, edited by H. Flap and J. Weesie. Utrecht, Netherlands: ISOR.

Mason, William M., G.Y. Wong, and Barbara Entwisle. 1983. "Contextual Analysis Through the Multilevel Linear Model." Pp. 72–103 in *Sociological Methodology 1983–1984*, edited by S. Leinhardt. San Francisco: Jossey-Bass.

Massey, Douglas S., and Nancy A. Denton. 1993. *American Apartheid: Segregation and the Making of the Underclass.* Cambridge, MA: Harvard University Press.

Matheron, Georges. 1971. *The Theory of Regionalized Variables and Its Applications.* Paris: Ecole Nationale Supérieure des Mines.

Matsueda, Ross L., and Karen Heimer. 1987. "Race, Family Structure, and Delinquency: A Test of Differential Association and Social Control Theories." *American Sociological Review* 52:826–40.

Mayer, Karl Ulrich, and Nancy Brandon Tuma. 1990. *Event History Analysis in Life Course Research.* Madison: University of Wisconsin Press.

McCullagh, Peter, and John A. Nelder. 1989. *Generalized Linear Models.* London: Chapman and Hall.

Morris, Carl N. 1983. "Parametric Empirical Bayes Inference: Theory and Applications" (with discussion). *Journal of the American Statistical Association* 78:47–65.

Morris, Martina. 1997. "Sexual Networks and HIV." *AIDS* 11:S209-16.

Morris, Martina, Annette D. Bernhardt, and Mark S. Handcock. 1994. "Economic Inequality: New Methods for New Trends." *American Sociological Review* 59:205–19.

Muthén, Bengt. 1983. "Latent Variable Structure Equation Modeling with Categorical Data." *Journal of Econometrics* 22:43–65.

———. 1997. "Latent Variable Modeling of Longitudinal and Multilevel Data." *Sociological Methodology* 27:453–80.

Nakao, Keiko, and Judith Treas. 1994. "Updating Occupational Prestige and Socioeconomic Scores: How the New Measures Measure Up." Pp. 1–72 in *Sociological Methodology 1994*, edited by P. V. Marsden. Cambridge, MA: Blackwell Publishers.

Nelder, John A., and R. W. M. Wedderburn. 1972. "Generalised Linear Models." *Journal of the Royal Statistical Society*, Ser. A, 135:370–84.

Padgett, John F., and C. K. Ansell. 1993. "Robust Action and the Rise of the Medici." *American Journal of Sociology* 98:1259–319.

Pearl, Judea. 1998. "Graphs, Causality, and Structural Equation Models." *Sociological Methods and Research* 27:226–84.

Petersen, Trond. 1991. "The Statistical Analysis of Event Histories." *Sociological Methods and Research* 19:270–323.

———. 1995. "Models for Interdependent Event-History Data: Specification and Estimation." Pp. 317–76 in *Sociological Methodology 1995*, edited by Peter V. Marsden. Cambridge, MA: Blackwell Publishers.

Poole, David, and Adrian E. Raftery. 2000. "Inference from Deterministic Simulation Models: The Bayesian Melding Approach." *Journal of the American Statistical Association* 95:1244–55.

Raftery, Adrian E. 1986a. "Choosing Models for Cross-Classifications." *American Sociological Review* 51:145–46.

———. 1986b. "A Note on Bayes Factors for Log-Linear Contingency Table Models with Vague Prior Information." *Journal of the Royal Statistical Society*, Ser. B, 48:249–50.

———. 1991. "Bayesian Model Selection and Gibbs Sampling in Covariance Structure Models." Working Paper 92-4, Center for Studies in Demography and Ecology, University of Washington.

———. 1995. "Bayesian Model Selection in Social Research" (with discussion). *Sociological Methodology* 25:111–93.

———. 1996. "Approximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear Models." *Biometrika* 83:251–66.

———. 1999. "Bayes Factors and BIC: Comment on 'A Critique of the Bayesian Information Criterion for Model Selection.'" *Sociological Methods and Research* 27:411–27.

Raftery, Adrian E., Geof H. Givens, and Judith E. Zeh. 1995. "Inference from a Deterministic Population Dynamics Model for Bowhead Whales" (with discussion). *Journal of the American Statistical Association* 90, 402–30.

Raftery, Adrian E., and Michael Hout. 1985. "Does Irish Education Approach the Meritocratic Ideal? A Logistic Analysis." *Economic and Social Review* 16:115–40.

Raftery, Adrian E., Steven M. Lewis, and Akbar Aghajanian. 1995. "Demand or Ideation? Evidence from the Iranian Marital Fertility Decline." *Demography* 32:159–82.

Ragin, Charles. 1987. *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*. Berkeley, CA: University of California Press.

Rao, J. N. K. 1996. "On Variance Estimation with Imputed Survey Data." *Journal of the American Statistical Association* 91:499–506.

Raudenbush, Stephen W., and Robert J. Sampson. 1999. "Ecometrics: Toward a Science of Assessing Ecological Settings, with Application to the Systematic Social Observation of Neighborhoods." *Sociological Methodology* 29:1–41.

Reiss, Albert J., Jr. 1971. "Systematic Observations of Natural Social Phenomena." *Sociological Methodology* 3:3–33.

Roberts, Carl W. 1997. "A Generic Semantic Grammar for Quantitative Text Analysis: Applications to East and West Berlin News Content from 1979." *Sociological Methodology*, 27, 89–130.

Roeder, Kathryn, G. S. Lynch, and Daniel S. Nagin. 1999. "Modeling Uncertainty in Latent Class Membership: A Case Study in Criminology." *Journal of the American Statistical Association* 94:766–76.

Rogoff, Nathalie. 1953. *Recent Trends in Occupational Mobility*. Glencoe, IL: Free Press.

Rubin, Donald, B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66:688–701.

———. 1977. "Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys." *Journal of the American Statistical Association* 72:538–43.

———. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

———. 1996. "Multiple Imputation After 18+ Years." *Journal of the American Statistical Association* 91:473–89.

Sampson, Robert J., and Stephen W. Raudenbush. 1999. "Systematic Social Observations of Public Spaces: A New Look at Neighborhood Disorder." *American Journal of Sociology* 105:603–51.

Sankoff, D., and Kruskal, J. B. 1983. *Time Warps, String Edits, and Macromolecules*. Reading, MA: Addison-Wesley.

Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.

Scheines, R., H. Hoijtink, and A. Boomsma. 1999. "Bayesian Estimation and Testing of Structural Equation Models." *Psychometrika* 64:37–52.

Schuessler, Karl F. 1980. "Quantitative Methodology in Sociology: The Last 25 Years." *American Behavioral Scientist* 23:835–60.

Schwarz, Gideon. 1978. "Estimating the Dimension of a Model." *Annals of Statistics* 6:461–64.

Singer, Burton, Carol D. Ryff, Deborah Carr, and W. J. Magee. 1998. "Linking Life Histories and Mental Health: A Person-Centered Strategy." *Sociological Methodology* 28:1–52.

Snijders, Tom A. B., and Roel J. Bosker. 1999. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Beverly Hills, CA: Sage.

Sobel, Michael E. 1981. "Diagonal Mobility Models: A Substantively Motivated Class of Designs for the Analysis of Mobility Effects." *American Sociological Review* 46:893–906.

———. 1985. "Social Mobility and Fertility Revisited: Some New Methods for the Analysis of Mobility Effects Hypothesis." *American Sociological Review* 50:699–712.

———. 1990. "Effect Analysis and Causation in Linear Structural Equation Models." *Psychometrika* 55:495–515.

———. 1994. "Causal Inference in Latent Variable Models." Pp. 3–35 in *Latent Variables Analysis: Applications for Developmental Research*, edited by A. von Eye and C. C. Clogg. Thousand Oaks, CA: Sage.

———. 1995. "Causal Inference in the Social and Behavioral Sciences." Pp. 1–38 in *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, edited by G. Arminger, C. C. Clogg, and M. E. Sobel. New York: Plenum.

———. 1997. "Measurement, Causation and Local Independence." Pp. 11–28 in *Latent Variable Modeling and Applications to Causality*, edited by M. Berkane. New York: Springer-Verlag.

———. 1998. "Causal Inference in Statistical Models of the Process of Socioeconomic Achievement: A Case Study." *Sociological Methods and Research* 27: 318–48.

———. 2000. "Causal Inference in the Social Sciences." *Journal of the American Statistical Association* 95:647–51.

Sobel, Michael E., Mark P. Becker, and Susan S. Minick. 1998. "Origin, Destination and Association in Occupational Mobility." *American Journal of Sociology* 104: 687–701.

Sorensen, Ann Marie. 1989. "Husbands' and Wives' Characteristics and Fertility Decisions: A Diagonal Mobility Model." *Demography* 26:125–35.

Spiegelhalter, David J., A. Philip Dawid, Steffan Lauritzen, and R. Cowell. 1993. "Bayesian Analysis in Expert Systems." *Statistical Science* 8:219–82.

Spirtes, Peter, Clark Glymour, and Richard Scheines. 1993. *Causation, Prediction and Search.* New York: Springer-Verlag.

Spirtes, Peter, Thomas S. Richardson, Christopher Meek, Richard Scheines, and Clark Glymour. 1998. "Using Path Diagrams as a Structural Equation Modeling Tool." *Sociological Methods and Research* 27:182–225.

Stern, R. D., and R. Coe. 1984. "A Model Fitting Analysis of Daily Rainfall Data" (with discussion). *Journal of the Royal Statistical Society*, Ser. A, 147:1–34.

Stovel, Katherine, M. Savage, and Peter S. Bearman. 1996. "Ascription into Achievement: Models of Career Systems at Lloyds Bank, 1890–1970." *American Journal of Sociology* 102:358–99.

Strang, David. 1991. "Adding Social Structure to Diffusion Models: An Event History Framework." *Sociological Methods and Research* 19:324–53.

Strang, David, and Nancy Brandon Tuma. 1993. "Spatial and Temporal Heterogeneity in Diffusion." *American Journal of Sociology* 99:614–39.

Thompson, Elizabeth A. 1998. "Inferring Gene Ancestry: Estimating Gene Descent." *International Statistical Review* 66:29–40.

Tobin, James. 1958. "Estimation of Relationships for Limited Dependent Variables." *Econometrika* 26:24–36.

Treiman, Donald J. 1977. *Occupational Prestige in Comparative Perspective.* New York: Academic Press.

Truett, J., Jerome Cornfield, and W. Kannel. 1967. "A Multivariate Analysis of the Risk of Coronary Heart Disease in Framingham." *Journal of Chronic Diseases* 20:511–24.

Tuma, Nancy Brandon. 1976. "Rewards, Resources, and Rates of Mobility: A Nonstationary Multivariate Stochastic Model." *American Sociological Review* 41:338–60.

Tuma, Nancy Brandon, and Michael T. Hannan. 1984. *Social Dynamics: Models and Methods*. Orlando, FL: Academic Press.

Udry, J. R., and Peter S. Bearman. 1998. "New Methods for New Research on Adolescent Sexual Behavior." Pp. 241–69 in *New Perspectives on Adolescent Risk Behavior*, edited by R. Jessor. Cambridge, England: Cambridge University Press.

Verma, T., and Judea Pearl. 1988. "Causal Networks: Semantics and Expressiveness." Pp. 352–59 in *Proceedings of the 4th Workshop on Uncertainty in Artificial Intelligence*. Mountain View, CA.

Warren, John R., Jennifer T. Sheridan, and Robert M. Hauser. 1998. "Choosing a Measure of Occupational Standing—How Useful Are Composite Measures in Analyses of Gender Inequality in Occupational Attainment?" *Sociological Methods and Research* 27:3–76.

Wasserman, Stanley, and K. Faust. 1994. *Social Network Analysis: Methods and Applications*, Cambridge, England: Cambridge University Press.

Wasserman, Stanley, and Philippa Pattison. 1996. "Logit Models and Logistic Regressions for Social Networks. 1. An Introduction to Markov Graphs and *p*," *Psychometrika* 61:401–25.

Weakliem, David L. 1992. "Does Social Mobility Affect Political Behavior?" *European Sociological Review* 8:153–65.

———. 1999. "A Critique of the Bayesian Information Criterion for Model Selection" (with discussion). *Sociological Methods and Research* 27:359–443.

Western, Bruce. 1996. "Vague Theory and Model Uncertainty in Macrosociology." *Sociological Methodology* 26:165–92.

Western, Bruce, and Simon Jackman. 1994. "Bayesian Inference for Comparative Research." *American Political Science Review* 88:412–23.

White, Harrison C. 1963. *An Anatomy of Kinship: Mathematical Models for Structures of Cumulated Roles*. Englewood Cliffs, NJ: Prentice-Hall.

Winship, Christopher, and Robert D. Mare. 1992. "Models for Sample Selection Bias." *Annual Review of Sociology* 18:327–50.

Wong, G. Y., and William M. Mason. 1985. "The Hierarchical Logistic Regression Model for Multilevel Analysis." *Journal of the American Statistical Association* 80:513–24.

Wright, Sewall. 1921. "Correlation and Causation." *Journal of Agricultural Research* 20:557–85.

Xie, Yu. 1992. "The Log-Multiplicative Layer Effect Model for Comparing Mobility Tables." *American Sociological Review* 57:380–95.

———. 1994. "Log-Multiplicative Models for Discrete-Time, Discrete-Covariate Event History Data," *Sociological Methodology*, 24, 301–40.

———. 2000. "Demography: Past, Present and Future." *Journal of the American Statistical Association* 95:670–73.

Xie, Yu, and Daniel Powers. 2000. *Statistical Methods for Categorical Data Models*. New York: Academic Press.

Yamaguchi, Kazuo. 1987. "Models for Comparing Mobility Tables: Towards Parsimony and Substance." *American Sociological Review* 52:482–94.

———. 1990. "Homophily and Social Distance in the Choice of Multiple Friends: An Analysis Based on Conditionally Symmetric Log-Bilinear Association Models." *Journal of the American Statistical Association* 85:356–66.

———. 1991. *Event History Analysis.* Newbury Park, CA: Sage.

———. 1992. "Accelerated Failure-Time Regression Models with a Regression Model of Surviving Fraction: An Application to the Analysis of 'Permanent Employment' in Japan," *Journal of the American Statistical Association* 87:284–92.

———. 1994. "Some Accelerated Failure-Time Models Derived from Diffusion Process Models: An Application to Diffusion Process Analysis." *Sociological Methodology* 24:267–301.

Yamaguchi, Kazuo, and Linda R. Ferguson. 1995. "The Stopping and Spacing of Childbirths and their Birth-History Predictors: Rational-Choice Theory and Event-History Analysis." *American Sociological Review* 60:272–98.

Yamaguchi, Kazuo, and Denise B. Kandel. 1998. "Parametric Event Sequence Analysis: Racial/Ethnic Differences in Patterns of Drug-Use Progression." *Journal of the American Statistical Association* 91:1388–99.