Mitchell, J. Clyde
1973 "Networks, norms, and institutions." Pp. 15–36 in J. Boissevain and J. Clyde Mitchell (eds.), Network Analysis: Studies in Human Interaction. The Hague: Mouton.
1979 "Networks, algorithms, and analysis." Pp. 425–51 in Paul Holland and Samuel Leinhardt (eds.), Perspectives on Social Network Research. New York: Academic Press.

Nadel, S. F.
1957 The Theory of Social Structure. London: Cohen & West.

White, Douglas R.
1982 "Rethinking the role concept: homomorphisms on social network." In Lin Freeman, Douglas White, and A. K. Romney (eds.), Research Methods in Social Network Analysis. Unpublished.

White, Douglas R. and Karl Reitz
Forthcoming "Graphs and semigroup homomorphisms on networks of relations." Social Networks.

White, Harrison C., Scott A. Boorman and Ronald L. Breiger
1976 "Social structure from multiple networks. I. Blockmodels of roles and positions." American Journal of Sociology 81:730–80.

Winship, Christopher
1976 "Roles and relations." Unpublished paper, Department of Sociology, Harvard University.

Winship, Christopher and Michael J. Mandel
Forthcoming "Roles and positions: a critique and extension of the blockmodelling approach." In Samuel Leinhardt (ed.), Sociological Methodology 1983/84. San Francisco: Jossey-Bass.

Wu, Lawrence L.
1980 "Roles structures in networks of trade and economic interdependence: a local blockmodel algebraic approach." Seniors Honors Thesis, Department of Sociology, Harvard University.
1982 "Local blockmodel algebras." Unpublished paper, Department of Sociology, Stanford University.

AN INTRODUCTION TO SAMPLE SELECTION BIAS IN SOCIOLOGICAL DATA*

RICHARD A. BERK
University of California, Santa Barbara

Sampling has long been central in discussions of sociological research methods. Yet, with few exceptions, recent developments on the nature of sampling bias have not filtered into sociological practice. This neglect represents a major oversight with potentially dramatic consequences since internal as well as external validity is threatened. In response, this paper undertakes a brief review of recent advances in the diagnosis of and corrections for "sample selection bias."

Sampling has long been central in discussions of sociological research methods. Yet, with a few exceptions (e.g., Tuma et al., 1979; Rossi et al., 1980; Berk et al., 1981), recent developments on the nature of sampling bias have not filtered into sociological practice. This neglect represents a major oversight with potentially dramatic consequences. More than external validity is threatened. Internal validity is equally vulnerable even if statements are made conditional upon the available data.

This paper undertakes a brief review of recent advances in the diagnosis of and corrections for "sample selection bias." Key points are illustrated with analyses taken from real data sets. Thus, the paper is no substitute for a careful reading of the primary source material and recent more lengthy, technical overviews. My goal is to direct the attention of the sociological community to a significant methodological problem while stressing major themes and intuitive reasoning.

WHAT'S THE PROBLEM?

Sample selection bias can be intuitively understood through the usual bivariate scatter plot interpreted within the framework of the general linear model. Given a fixed regressor (gener-

* Direct all correspondence to: Richard A. Berk, Department of Sociology, University of California, Santa Barbara, CA 93106.

The research reported in this paper was supported by a grant from the National Institute of Justice (grant No. 80–IJ–CX–0037). I am also grateful for the help in data collection provided by Anthony Shih and Jimy Sanders. Finally, Karl Schuessler, Kenneth Land, and Phyllis Newton provided helpful comments on an earlier draft of the paper.

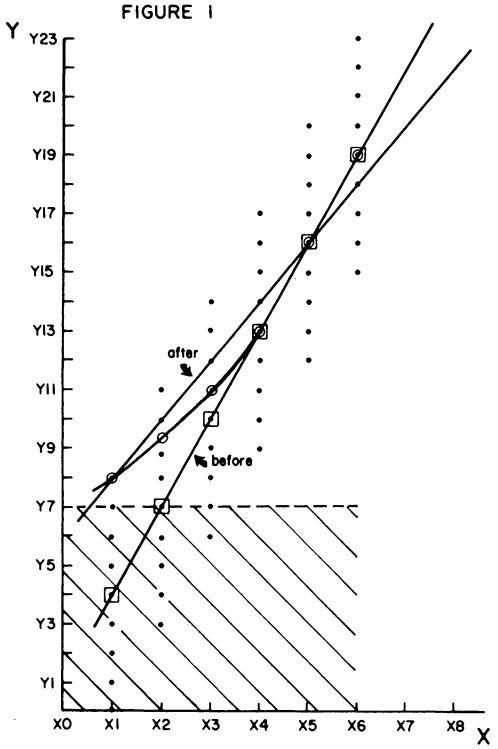
alizations to stochastic regressors are easily accomplished, e.g., Pindyck and Rubinfeld, 1981:274–78), one assumes a linear relationship between an exogenous and an endogenous variable. One also assumes that the endogenous variable is affected additively by a disturbance (error) term characterized by an expected value of zero for each value of the exogenous variable.[1] If these two assumptions are met, the disturbance term is uncorrelated with the exogenous variable, which guarantees unbiased least squares estimates of the slope and intercept. Other assumptions about the disturbance term that are typically made need not concern us.[2]

Figure 1 is a scatter plot for an endogenous variable Y and an exogenous variable X. Assume the data are a simple random sample from some population of interest, that in this population the linear form is correct, and that for each value of X the mean of the disturbances is zero. Implied is that the regression line passes through the expected value of Y for each value of X. In Figure 1 these expected values are represented by boxes, and the regression line is labeled "before."

In Figure 1, suppose that observations with values on Y equal to or less than Y7 cannot be obtained. For example, suppose that Y is a measure of the seriousness of incidents of wife battery, and that police only make an arrest in such incidents if the dispute exceeds some level of seriousness (Berk et al., 1983). Then, if one's data are taken exclusively from police arrest reports, less serious incidents will be systematically underrepresented. In Figure 1 observations in the shaded area are missing.

For low values of X in Figure 1 the new expected values are represented by circles. Thus, for all observations with X equal to X1, the expected value of Y has shifted from Y4 to Y8. Likewise, for all observations with X equal to X2, the expected value of Y has shifted from Y7 to Y9.5. As X increases, the size of the shift is reduced until by X4, the new and old expected values are virtually identical.

FIGURE 1



The new expected values for Y means that the original regression line no longer fits the data. The relationship between X and Y is no longer linear; the slope becomes steeper as X increases (up to X4). Consequently, *any* attempt to fit a straight line will produce a specification error. Basically, one is using the wrong functional form. In Figure 1 the second regression line labeled "after" shows the result that might materialize. Compared to the true relationship, the estimated relationship has been attenuated.

What are the implications? First, external validity has been undermined. The regression line estimated from the scatter plot in Figure 1 will systematically underestimate the slope of the population regression line. If X is the number of prior wife battery incidents, the estimated causal effect of such priors on the seriousness of the immediate incident will be substantially smaller than the causal effect in the population. Excluding less serious incidents

[1] Actually, one assumes that for each observation, the expectation of the disturbance term is zero; this implies that the expectations for each value of the exogenous variable are zero. The assumption of linearity allows for nonlinear relationships that can be transformed into linear ones (e.g., Pindyck and Rubinfeld, 1981:107–110). With time series data, one sometimes makes a distinction between exogenous variables and predetermined variables. All regressors are predetermined, including lagged values of the endogenous variable. However, lagged values of the endogenous variable are not exogenous. A further discussion of such issues can be found in Engle et al. (1983).

[2] In order to obtain efficient estimates of the regression coefficients and unbiased estimates of their standard errors, one must assume that the disturbances are uncorrelated with one another and that all have the same variance. Then in "small" samples, one must assume for significance tests that the disturbances are normally distributed. Asymptotically, the normality assumption is unnecessary. Discussion of the assumptions for least squares procedures can be found in virtually any econometrics text.

attenuates the causal effect in this instance. Clearly, one should not try to generalize from the sample in Figure 1 to all incidents of wife battery. Such problems are well understood by most sociologists.

Second, and not commonly recognized, internal validity is also jeopardized *even if one is prepared to make causal inferences to a population of less serious battery incidents*. In Figure 1, for low values of X, the regression line falls on or above the expected values, while for high values of X, the regression line falls on or below the expected values. For low values of X, therefore, negative disturbances will predominate, while for high values of X, positive disturbances will predominate. This implies that X will be positively correlated with the disturbance term. As a result, least squares estimates of the slope and intercept will be biased (and inconsistent as well), even if one is only interested in the causal relationship between the seriousness of the incident and the number of priors for the subset of more serious incidents. Put another way, effects of the exogenous variable and the disturbance term are confounded, and causal effects are attributed to X that are really a product of random perturbations.

The confounding of X and the disturbance term follows in this example *even if one's sole concern is with more serious wife battery incidents*. One cannot dismiss the problem by claiming interest only in the nonrandom subset of cases represented by the sample at hand. By excluding some observations in a systematic manner, one has inadvertantly introduced the need for an additional regressor that the usual least squares procedures ignore (Heckman, 1976, 1979); in effect, one has produced the traditional specification error that results when an omitted regressor is correlated with an included regressor (e.g., Kmenta, 1971:392–95).

Figures 3 through 5 present in schematic fashion other examples of outcomes obtained when segments of some population cannot be observed. Figure 2 is a new representation of Figure 1 and serves as a benchmark.

Suppose in Figure 3 that Y is income and X is education and that the sample only includes individuals with income below the poverty line. The estimated regression line is again biased downward with both external validity and internal validity weakened. One cannot generalize the estimated causal relationship to all adults nor is the relationship between education and income properly represented, *even for individuals with incomes below the poverty line*.

Both Figure 2 and 3 depict exclusion through a threshold for the endogenous variable under scrutiny. Goldberger (1981), borrowing from Lord and Novick (1968), has called this manner of selection "explicit." Alternatively, one might use the term "direct" for reasons that will be apparent shortly.[3]

Figure 4 shows a more complicated selection process. The lower right-hand section of the scatter plot has been eliminated, but not in a way that reflects a single threshold on Y. How might this happen? Suppose that Y is the amount of money spent on medical care, and X is the amount a person smokes. Also suppose that people who smoke more are more likely to have fatal illnesses, other things being equal. Clearly, one cannot observe the amount of money spent on medical care for individuals who are no longer alive.

It is important to stress that no threshold is defined in terms of medical costs or even the amount of smoking. Rather, the threshold involves a new variable, physiological viability, that, for purposes of illustration, has been assumed not to play a role in the relationship of interest (i.e., the effect of smoking on medical costs). When physiological viability falls below the threshold of death, the case is excluded. Goldberger, again drawing from Lord and Novick, has called such selection processes "incidental." Alternatively, one might use the term "indirect."

As before, both external validity and internal validity are jeopardized. Once again, the exclusion of a nonrandom subset of observations introduces a nonlinear relationship between X and Y. When, in this instance, a straight line is fitted, the estimated causal relationship is inflated, and effects attributed to X include the impact of the disturbance term.

Consider a second example. Suppose that Y is length of time retail stores remain in business, and X is amount of capital the stores had when they opened. Suppose also that in 1970 one obtains a random sample of retail stores just opening for business and that data are collected until 1980. However, not all stores fail in the ten-year interval; for some fraction of the cases the time to failure cannot be observed. Such incidental selection is called right-hand censoring in the failure time literature (e.g., Lawless, 1982; Tuma, 1982) and can lead to distorted scatter plots as in Figure 4. Less common, left-hand censoring is also possible

[3] Explicit selection can be generalized so that the threshold is not a constant (Goldberger, 1981), but the generalization has not had a substantial impact on empirical work. There seems no need, therefore, to complicate the discussion with variable thresholds.

FIGURE 2

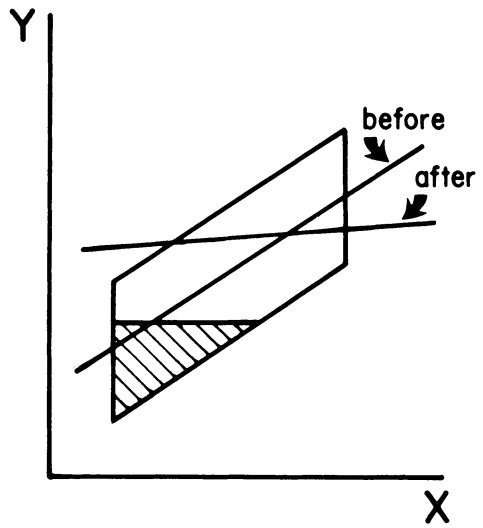


FIGURE 3

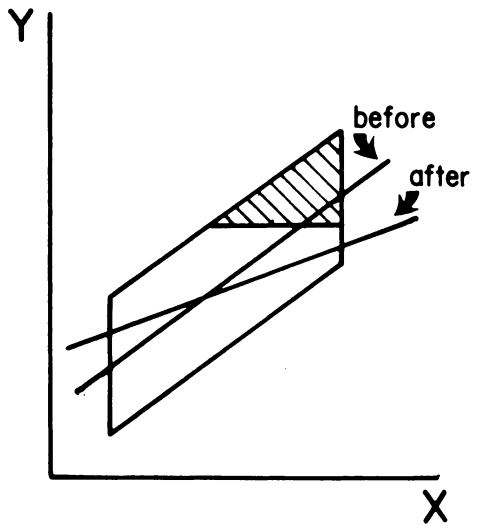


FIGURE 4

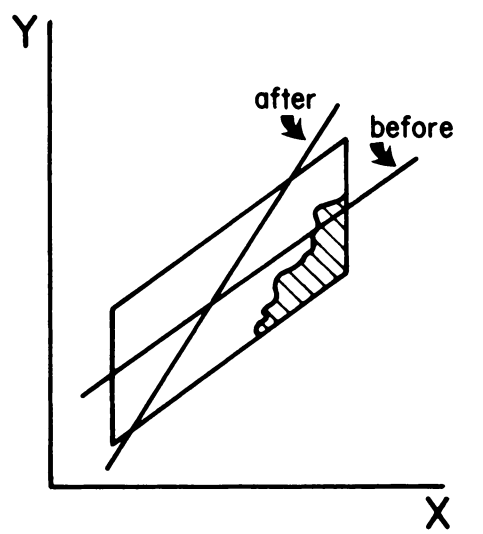
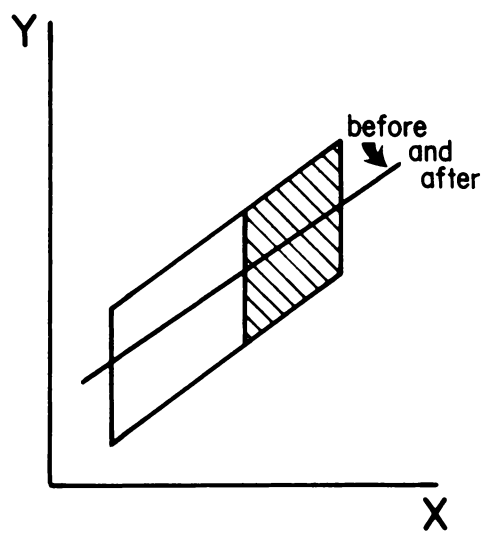


FIGURE 5



(i.e., the data collection begins after some units have failed).

Figure 5 shows a pattern in which a threshold for exclusion is defined for the exogenous variable. Suppose that people of all incomes are included, but people with greater than a high school education are not. If the relationship between education and income is really linear across the full range of educational levels, external validity and internal validity are unscathed.[4] One can generalize to a population that includes individuals with more than a high school education and estimate the effect of education in an unbiased manner.

There are, thus, three initial lessons to be

[4] The danger is that by excluding observations one has a smaller sample and perhaps less variation in X. Both reduce one's statistical power; standard errors will be increased.

learned. First, if potential observations from some population of interest are excluded from a sample on a nonrandom basis, one risks sample selection bias. Nonrandom exclusion of certain observations can be caused by data collection procedures or by processes inherent in the phenomena under study. For example, skip patterns are meant to weed out nonrandom subsets of respondents for whom some questions do not apply. Such procedures risk sample selection bias when the remaining (nonrandom) observations are analyzed. In this situation, a researcher's data collection procedures recapitulate nonrandom selection in the social world. The general point is that the prospect for sample selection bias is pervasive in sociological data. Circumstances under which the prospect becomes a reality will be addressed below.

Second, it is difficult to anticipate whether the biased regression estimates overstate or understate the true causal effects. The direction and size of the bias depends in the bivariate case on the number and location of observations that are excluded; the situation is enormously more complicated in multivariate models. When sample selection bias is present, one is essentially flying blind. One is faced with the same kinds of problems one finds in multiple regression analyses with conventional specification or measurement errors. Only in special cases[5] can the direction of the distortions be known.

Third, the problems caused by nonrandom exclusion of certain observations are manifested in the expected values of the endogenous variable. When the usual linear form is fit to the data, the expected values of the disturbances for each value of X are no longer zero. The bad news is that the disturbances are then correlated with the exogenous variable. The good news is that in the nonlinear form lies a potential solution.

A MORE FORMAL STATEMENT OF THE PROBLEM

The social science literature contains several formal introductions to sample selection bias (e.g., Heckman, 1976; Goldberger, 1981) and several textbook-level discussions (e.g., Judge et al., 1980: ch. 14; Berk and Ray, 1982). Probably the best known and most accessible formulation is by Heckman (1979). I have drawn heavily on his exposition.[6]

Consider a random sample of I observations with two equations of interest:

$$Y_{1i} = X_{1i}\beta_1 + U_{1i} \quad (1a)$$

$$Y_{2i} = X_{2i}\beta_2 + U_{2i} \quad (i = 1, \ldots, I), \quad (1b)$$

where each X is a vector of exogenous variables which may, or may not, be the same, and the betas are vectors of conformable regression coefficients. In both equations, the expected values of the disturbances are taken to be zero, which implies that both equations are properly specified. More generally, each equation by itself is assumed to meet the usual assumptions for ordinary least squares. Across equations, however, the disturbances are correlated and are assumed to behave as if drawn from a bivariate normal distribution. Thus, equations 1a and 1b represent a pair of seemingly unrelated equations (e.g., Pindyck and Rubinfeld, 1981:323–24). One has nothing more than a pair of regression equations with correlated disturbances.

Suppose that on sociological grounds one cares about the first equation; equation 1a can be thought of as the "substantive equation." However, one can only observe the endogenous variable in that equation if the endogenous variable in the second equation exceeds (or does not exceed) some threshold. The second equation can be called the "selection equation."

To make this more concrete, suppose that the first equation is a causal model of the length of prison sentences given to convicted felons. Yet, convictions can only result if the strength of the evidence implies guilt beyond a reasonable doubt; one can think of the second equation as a causal model for the strength of evidence. Individuals for whom reasonable doubt exists are excluded from sentencing. Since the same parties are involved in both the determination of guilt and the determination of sentence length, the disturbances in the two equations are plausibly correlated. That is, random perturbations (e.g., how aggressive the prosecutor is) will simultaneously affect both endogenous variables.

[5] Goldberger (1981) discusses the situations in which the direction and size of the bias can be determined. If one can assume the data come from a multivariate normal distribution, then in the case of explicit selection, all regression coefficients are attenuated. For incidental selection, even under the assumption of multivariate normality, the direction and size of the bias cannot be determined.

[6] The problem has a long history. Pearson and Lee (1908) wrestled with truncated distributions, the econometrics community was first introduced to the problem of explicit selection by James Tobin (1958), and biometricians have worried about left-hand and right-hand censoring at least as long (e.g., Lawless, 1982:34–44).

Equations 2a and 2b show the results of the selection process.

$$E(Y_{1i} \mid X_{1i}, Y_{2i} \geq 0) = X_{1i}\beta_1 + \frac{\sigma_{12}}{(\sigma_{22})^{1/2}}\lambda_i \quad (2a)$$

$$E(Y_{2i} \mid X_{2i}, Y_{2i} \geq 0) = X_{2i}\beta_2 + \frac{\sigma_{22}}{(\sigma_{22})^{1/2}}\lambda_i \quad (2b)$$

For equation 2a the conditional expectation of the endogenous variable is equal to the expected value of the original substantive equation (1a) plus a new term. For equation 2b the conditional expectation of the endogenous variable is equal to the expected value of the original selection equation (1b) plus a somewhat different new term. Focusing first on the substantive equation 2a (e.g., the equation for sentence length), the new term can be divided into two parts. The first part is the ratio of the covariance between the disturbances in equations 1a and 1b to the standard deviation of the disturbances in equation 2a. The ratio, therefore, serves as a regression coefficient; if the covariance between the two disturbances is zero, *the extra term disappears*. If the disturbances are uncorrelated, the usual least squares procedures will suffice.

The meaning of the second component can be understood through the following equations:

$$\lambda_i = \frac{f(z_i)}{1 - F(z_i)} \quad (3)$$

$$z_i = -\frac{X_{2i}\beta_2}{(\sigma_{22})^{1/2}} \quad (4)$$

The z in equation 4 is the negative of the predicted value from a probit equation in which one models the likelihood that in the selection equation (1b) the threshold will be equaled or exceeded. In our example, the probit equation models the likelihood that a conviction will occur. However, since the predicted value is multiplied by -1, one is ultimately capturing the likelihood that a conviction will not occur; the issue is really which cases will be *excluded*.

More specifically, the predicted value from a probit equation is a normally distributed, random variable with a mean of zero and a standard deviation of 1.0. The negative of this random variable is then used in equation 3, where the numerator is the variable's density, and the denominator is 1.0 minus the variable's (cumulative) distribution. The ratio is called the hazard rate, which represents for each observation the instantaneous probability of being excluded from the sample conditional upon being in the pool at risk (Tuma, 1982:8–10). The larger the hazard rate, the greater the likelihood that the observation will be discarded.

Equally important, the hazard rate captures the expected values of the disturbances in the substantive equation after the nonrandom selection has occurred. It was precisely these expected values that are the source of the biased estimates. By including the hazard rate as an additional variable, one is necessarily controlling for these nonzero expectations. Alternatively stated, the deviations of the expected values from the regression line result from an omitted variable that has now been included. The key, then, to consistent parameter estimates is to construct a hazard rate for each observation. And it cannot be overemphasized that it is the selection process that *introduces* the need for a new variable.

Turning to equation 2b, the hazard rate is constructed from the same equation in which it is then used; the distinction between the selection and substantive equations disappears. Referring back to our earlier terminology, the two-equation model (equations 2a and 2b) represents incidental or indirect selection. The one-equation model (equation 2b) represents explicit or direct selection. The latter is also known as a Tobit Model (Tobin, 1958).[7]

To summarize, whenever one has a nonrandom sample, the potential for sample selection bias exists. Examples are easy to construct. Studies of classroom performance of college students rest on the nonrandom subset of students admitted and remaining in school. Studies of marital satisfaction are based on the nonrandom subset of individuals married when the data are collected. Studies of worker productivity are limited to the employed. And, potential problems are complicated by inadequate response rates.

Alternatively stated, the difficulty is that one risks confounding the substantive phenomenon of interest with the selection process. The impact of a mother's level of education on a child's college grade point average may be

[7] When the selection process eliminates observations solely for the endogenous variable, one commonly speaks of censoring. When observations are missing in the exogenous variables as well, one commonly speaks of truncation (Heckman, 1976:478). Here, only censored samples are considered. Truncation causes far more serious difficulties that are well beyond the scope of this paper. An introduction to the issues and a good bibliography can be found in Berk and Ray (1982). It is also important not to confuse sample selection censoring or truncation with legitimately bounded endogenous variables where no observations are lost. For example, analyses of some kinds of survey questions must respond to ceiling and floor effects and, in a sense, these effects truncate the endogenous variable. However, floor and ceiling effects imply a nonlinear functional form (e.g., a logistic) and not a failure to observe certain values on the endogenous variable.

confounded with its impact on the child's likelihood of getting into college. The impact of a husband's income on the amount of leisure time a couple shares may be confounded with its impact on the likelihood that the couple will be married at all. The impact of seniority on output per hour may be confounded with its impact on the likelihood of being employed. Finally, the impact of a respondent's race on any of these phenomena may be confounded with its impact on the likelihood of responding to a questionnaire.

There is also the problem of infinite regress. Even if one has a random sample from a defined population, that population is almost certainly a nonrandom subset from a more general population. Suppose one has a random sample of all felony arrests in a given state in a given year. The random sample of felony arrests is a nonrandom sample of all reported felonies in that state in a given year. The reported felonies are a nonrandom sample of actual felonies committed. The felony arrests in a given state are also not a random sample of felonies in all states. In principle, therefore, there exists an almost infinite regress for any data set in which at some point sample selection bias becomes a *potential* problem. As for traditional specification errors and measurement errors, the question is not typically whether one has biased (or even consistent) estimates.[8] The question is whether the bias is small enough to be safely ignored.

Given the almost universal potential for sample selection bias, the critical issue becomes when that bias is likely to materialize. The key lies in the correlation between the disturbances for the substantive and selection processes. Under explicit selection, the substantive and selection processes are captured in a single equation. The two disturbance terms are, therefore, identical and correlate perfectly. Thus, *any* nonrandom (explicit) selection produces biased and inconsistent estimates of the regression coefficients, with the bias a function of the proportion of the sample excluded. If one is prepared to assume that the data (exogenous and endogenous variables) are drawn from a multivariate normal distribution, the bias is proportional to the probability of exclusion (Goldberger, 1981). And the probability can be estimated from the proportion of cases for which no observations on the endogenous variable are available. Explicit selection seems to be relatively rare in sociological data.

The situation for incidental selection is more complicated. One rarely knows much about the likely sign and magnitude of the correlation between the disturbances. Perhaps the easiest case is found when one can point to an obvious variable omitted from both the substantive and selection equations that is also *uncorrelated with the regressors included*. The omitted variable will cause the disturbances to be correlated, but since the omitted variable is uncorrelated with the included regressors, the equations (prior to sample selection) are properly specified (under the usual definition of specification error).

For example, one might be interested in victimization from natural disasters such as tornadoes, floods, and earthquakes (Rossi et al., 1982). Suppose that questionnaires are given to a random sample of adults and that response rates are virtually 100 percent. In an analysis of the amount of damage done in "the most recent" disaster, a large number of respondents would have nothing to report. Indeed, skip patterns in the questionnaire are designed to spare them from such items.

Almost regardless of how one conceives the substantive and selection processes, the severity of the natural disaster to which respondents were exposed (from no experience to a devastating one) should affect the likelihood of reporting a firsthand experience and also the amount of damage that resulted. However, if no external measure of disaster severity is available, no external measure of severity can be included in either equation. Should that omitted variable be correlated with regressors that are included, one has the traditional omitted variable specification error. If, however, one can argue that the severity of the natural disaster is probably uncorrelated with the included regressors, one can alternatively assert that sample selection bias will be present when data from the subset of disaster victims are analyzed. Given the processes that determine the location and magnitude of tornadoes, for example, the arguments for sample selection bias (rather than traditional omitted variable bias) may well be plausible. For example, it is unlikely that the probability of damage from a tornado is related to education, income, or attitudes toward risk.

In most sociological research, the issues are muddier. One must first justify the model specifications for the substantive and selection equations (no small feat) and then carefully address whether the disturbances are likely to be correlated. There are probably grounds for concern when the substantive and selection processes unfold with the same actors, and/or in the same physical locations, and/or at about the same time. Under these conditions, random perturbations will have a significant opportunity to affect jointly the selection and

[8] Randomized experiments come the closest to eliminating such problems.

substantive outcomes. The sentencing example is surely a good illustration. Studies of the wages earned by women are among the best known examples in the economics literature. One can only observe wages for women who are employed, and employed women are a nonrandom subset of all women. Moreover, random perturbations are likely to affect simultaneously both the probability of getting a job and wages once the job begins (Heckman, 1980). More generally, however, the social science community still has very limited experience with the sample selection problem, and there are as yet no compelling guidelines.

APPLICATIONS

Examples of corrections for explicit selection are readily found elsewhere, often under the rubric of Tobit Models (Tobin, 1958; Greene, 1981; Berk et al., 1983). In the pages ahead, analyses will be presented in which incidental selection is at issue.

In the analyses to be discussed shortly, the following steps are followed:

1. A probit model of the selection process is estimated with the dummy endogenous variable coded "0" when the observation on the substantive endogenous variable is missing and "1" when it was present.
2. The predicted values from the probit equation are saved. These predicted values represent a random, normal variable.
3. From the predicted values, the hazard rate is constructed. The predicted values are first multiplied by -1.0, and the density and distribution values calculated. The results are plugged into equation 3.
4. The hazard rate is then treated as a new variable and included in any substantive equations.
5. The bulk of the substantive analyses are done with ordinary least squares, although spot checking with other procedures (e.g., generalized least squares) is also undertaken.

The data are from a study of citizen opinions of various parts of the criminal justice system. For each of four county criminal justice agencies (a Police Department, The Office of the Court Administrator, The Public Defender's Office, and a Victim/Witness Assistance Program in the District Attorney's Office), self-administered questionnaires were mailed to random samples of individuals shortly after these individuals had an encounter with the agency in question. Here, we will rely exclusively on material from people who were called for jury duty. Overall, the problem was an effort to determine if accurate and cost-effective ways could be developed to provide rapid citizen feedback on the performance of the criminal justice system (Berk and Shih, 1982).

We anticipated low response rates. Therefore, we collected from official records considerable information on *all* prospective respondents, expecting to model failures to return the questionnaire.

Table 1 shows the three selection equations for nonresponse. The results on the far right derive from a probit model which rests on what we have been assuming so far: the two disturbances are bivariate normal. If one is prepared to assume that the disturbances are bivariate logistic, then the selection equation should be logistic (Ray et al., 1980). Finally, if one is prepared to assume that the disturbances in the selection equation follow a rectangular distribution and that the disturbances in the substantive equation are a linear function of the disturbances in the selection equation, the linear probability model may be used to model selection (Olsen, 1980b).

The probit approach is by far the most popular, and we will continue to rely on it. However, there is some concern in the literature about what happens if bivariate normality is violated, including what the appropriate options may be (Olsen, 1980a,b; Greene 1981; Arabmazar and Schmidt, 1982). The three sets of results are presented to stress that there are options to the assumption of bivariate normality, that these options are easy to implement, and to consider whether in this instance the results depend on the option chosen.

Five conclusions follow.[9] First, the response rate is nearly 70 percent, which is certainly respectable by social science standards. Thus, there may be too few observations excluded to introduce serious selection bias.

Second, using the full sample of 498, none of the three equations is very successful at explaining nonresponse. All are able to account for 5 percent of the variance. This may result from the omission of important exogenous variables or from near random patterns of nonresponse. If the former, proper corrections may not be feasible. If the latter, the hazard rate to be constructed will have little variance and will be unlikely to have a statistically significant regression coefficient in the substan-

[9] The coding conventions reported at the bottom of Table 1 follow from the derivation of the "hazard rate" for each of the three models. For all three, the goal is to construct a variable that captures the likelihood of exclusion from the sample (i.e., nonresponse). For the linear and logistic, this is accomplished in the way nonresponse is initially coded. As we pointed out earlier, for the probit form this is accomplished later when the new variable is constructed.

Table 1. Selection Equation for Non-Response (Response Rate = 69% of 498 Cases)

Variable	Linear[a]		Logistic[b]		Probit[c]	
	Coeff.	t-Value	Coeff.	t-Value	Coeff.	t-Value
Intercept	0.650	5.48	0.944	1.56	−0.546	1.44
Female Respondent (dummy)	−0.030	−0.72	−0.156	−0.77	0.096	0.78
Age (years)	−0.005	−3.44	−0.025	−3.49	0.015	3.26
Age "missing" (dummy)	0.102	1.38	0.507	1.47	−0.311	−1.54
Respondent Employed (dummy)	0.020	0.44	0.113	0.50	−0.079	−0.59
Respondent Served on Jury (dummy)	−0.107	−1.44	−0.545	−1.46	0.275	1.21
Served × Criminal Trial (dummy)	0.023	0.39	0.116	0.37	−0.036	0.20
Served × Length of Trial (dummy)	0.004	0.54	0.022	0.60	−0.009	−0.43
Served × Defendant Won (dummy)	0.077	1.07	0.371	1.07	−0.21	−1.00
Length of Jury Selection (dummy)	−0.088	−1.25	−0.494	−1.32	0.315	1.35
	R^2 = .05		D = .05		R^2 = .05	
	F = 2.59		χ^2 = 23.87		F = 2.74	
	P = <.01		P = <.01		P = <.05	

Descriptive Statistics for Instruments

	N	Mean	Standard Deviation	Minimum	Maximum
Linear	498	−0.70	0.10	−0.99	0.50
Logistic	498	0.30	0.10	0.07	0.53
Probit	498	0.48	0.14	0.12	0.81

[a] 0 = replied, 1 = did not reply.
[b] 0 = replied, 1 = did not reply.
[c] 1 = replied, 0 = did not reply.

tive equation. Near the bottom of the table are shown descriptive statistics for the "hazard rate" variables (a kind of instrumental variable) constructed from the three equations.[10]

Third, keeping in mind the coding conventions listed at the bottom of the table (see footnote 9), the story across the three equations is virtually identical. Perhaps the easiest way to compare across the equations is to examine the three t-values for each parameter estimate. Alternatively, there are approximate transformations between the three sets of coefficients (Amemiya, 1981). For example, if each of the regression coefficients in the probit model is multiplied by .40, approximations of the linear coefficients follow.

Fourth, only one variable has a statistically significant effect on the likelihood of nonresponse at conventional levels. Interpreting the linear coefficient, for each ten years of age the probability of nonresponse decreases by 5 percent. There is also a hint that if the respondent was subjected to a more lengthy jury selection process or was selected to serve as a juror, the likelihood of nonresponse declines. Perhaps greater involvement at the courthouse leads to greater involvement in the questionnaire. Finally, for about 10 percent of the cases age was not available from the official records. For these individuals, the mean was inserted. To control for some distortions that might result, a dummy variable was included, coded "1" for those cases. However, since the official measure of age was routinely obtained from very short questionnaires mailed to all prospective jurors by the Jury Commissioner, we suspected that individuals who did not cooperate fully with the Jury Commissioner would be less cooperative with us. There is a bit of evidence that this is true. Still, the results in Table 1 are not especially instructive.[11]

Fifth, all three "hazard rates" were constructed and correlations were calculated among them. For these data, the lowest correlation is .98. Clearly, it would not matter (and in fact does not matter) which version of the "hazard rate" is used. There is, however, no reason to believe that this is a general result

[10] The "hazard rate" from the linear probability model is equal to the predicted probability of nonresponse minus 1.0. The "hazard rate" from the logit model is simply the predicted probability of nonresponse.

[11] It is possible to find selection effects in one's substantive equation, even if one cannot find systematic selection effects in the selection equation itself (Heckman, 1979:155). However, only the intercept in the substantive equation is altered.

Table 2. Ordinary Least Squares Analysis of Overall Dissatisfaction

"All in all, how would you rate your experience of being called for jury duty?"			
Very Satisfied = 3 39.9% (135)	Somewhat Satisfied = 2 37.6% (127)	Somewhat Dissatisfied = 1 15.7% (53)	Very Dissatisfied = 0 6.8% (23)

	Uncorrected		Probit Correction	
Variable	Coeff.	t-Value	Coeff.	t-Value
Intercept	1.47	5.31	2.26	5.91
Hazard Rate	—	—	−1.26	−2.97
Female	0.28	2.91	0.26	2.67
Employed	0.03	0.32	0.15	1.44
White	0.16	1.18	0.12	0.92
Served	0.37	2.07	0.21	1.13
Served × Criminal Trial	0.11	0.73	0.10	0.70
Served × Length of Trial	−0.01	−0.68	−0.00*	0.25
Served × Defendant Won	0.07	0.40	0.21	1.20
Length of Jury Selection	−0.17	−1.07	−0.33	−1.98
1st Time Called	0.16	1.57	0.20	1.99
# Days Notice	0.11	1.85	0.12	1.98
Does Not Drive	0.05	0.31	0.07	0.48
	R^2 = .10 F = 3.41 P = < .001		R^2 = .13 F = 3.94 P = < .001	

* Negative, but smaller than 0.00.

and may be a consequence of the small amount of variance explained in each of the three selection equations; all three constructed "hazard rates" may be insufficiently variable to reveal properly their different forms.

Table 2 shows the results for one of the questionnaire items. Among those who returned the questionnaire, nearly 80 percent were at least somewhat satisfied with the experience. For most of the other questions, similar sentiment was expressed (Berk and Shih, 1982).

Turning to the multivariate equations, the left-hand side shows the usual least squares coefficients. The results on the right-hand side have been corrected through the addition of the hazard rate instrument. Perhaps the most important message is that the uncorrected and corrected results differ substantially. With the addition of the hazard rate, 3 percent more variance is explained, and the regression coefficient for the hazard rate is statistically significant at well beyond conventional levels ($t=-2.97$). The sign of the regression coefficient indicates that individuals who are less likely to return the questionnaire are more critical of the jury experience; complainers are less inclined to respond.

More important, the uncorrected equations include *false positives and false negatives*. Statistically significant coefficients would have been overlooked for the length of the jury selection, whether the respondent had previously been called for jury duty, and the number of days' notice given (assuming a two-tailed test). After corrections are made,[12] respondents are more positive if they are first timers, if they are given more notice, and the time taken for jury selection is shorter. All three effects have important policy implications (Berk and Shih, 1982) that would have been lost had corrections for sample selection bias not been undertaken. Note also that the relative importance of the jury selection variable has been substantially altered.

In the uncorrected equation, there is one false positive; individuals who served on a jury are incorrectly deemed more positive. In other words, one would have falsely concluded that serving on a jury by itself led to more favorable assessments.

Finally, only one causal effect holds in both the corrected and uncorrected equations. Female respondents are in both instances more complimentary. This was a general result over a wide variety of items.

There is, however, at least one important ambiguity. With the correction, there is a substantial change in the intercept, perhaps implying an increase in the mean of the endogenous variable. That is, the change in the intercept suggests that the original regression

[12] The corrected results are nothing more than ordinary least squares with the hazard rate included. Technically, generalized least squares is superior, but by using procedures outlined by Heckman (1976:483) little of interest changes. Still better would have been maximum likelihood procedures, but no software was available.

hyperplane was artifactually low (i.e., too critical). Unfortunately, the hazard rate is very collinear with the intercept; with the correction, the standard error of the intercept increases by about 50 percent. When a confidence interval is placed around the corrected intercept, it is no longer clear that a dramatic shift has occurred.[13] In short, we suspect that in this instance the selection bias primarily affects the regression coefficients.

The substantive and policy-related implications of these results are discussed elsewhere (Berk and Shih, 1982). Perhaps the major conclusion is that for the jury data, sample selection bias stemming from nonresponse is pervasive *even with a response rate of nearly 70 percent*. People who are perhaps alienated from society and its institutions are less likely to return the questionnaire and more critical of the jury experience.

Here, the emphasis is on method, and the major conclusion is superficially straightforward: the hazard rate correction clearly makes a substantial difference. One problem, however, is that there is really no way of knowing whether the correction is responding to a real sample selection bias or a pseudo-bias produced by preselection specification errors in the substantive and selection equations. It is easy to think of interesting variables that were unavailable for respondents who did not return the questionnaire (e.g., race) and interesting variables that were unavailable altogether (e.g., education). Yet, it is not obvious for the population of registered voters (from which jurors are chosen) that important omitted variables are likely to exist that are correlated with the variables included. Another problem is sample selection bias we may have overlooked. In particular, the population of registered voters is hardly a random subset of all county residents. This resurrects the earlier point about infinite regress.

CONCLUSIONS AND GENERALIZATIONS

The potential for sample selection bias exists whenever one is working with a nonrandom subset of some population. While our application was necessarily limited to selection problems within a survey framework, other data forms are hardly immune: administrative records (e.g., from schools), observational material (e.g., check lists of activities from "ride-alongs" with police officers), business documents (e.g., membership on the Board of Directors), and the like. Much like specification error and measurement error, the potential for bias is virtually universal.

Both internal and external validity are implicated. There is no escape by limiting one's causal conclusions to the population from which the nonrandom sample was drawn (or even the sample itself). Most sociologists recognize the threat to external validity; the threat to internal validity has generally been overlooked.

When considering whether potential sample selection bias is likely to be realized, the initial step is to formulate a theoretical model of the selection process. One needs a *theory of selection*. Without a theory, it is difficult to draw even preliminary inferences about the nature of the problem and impossible to choose how best to implement sample selection corrections. Unfortunately, while considerable effort has been devoted to documenting sampling biases within traditional survey sample approaches (e.g., Dillman, 1978), we are a very long way from a formal theory. Perhaps more important, other ways in which sample selection may be introduced are only beginning to be examined.

Given the potential for sample selection bias, the key lies in the correlation between disturbances of the substantive and selection equations.[14] When the correlation is zero, the potential for sample selection bias does not materialize. When the correlation is small or, in the case of explicit selection, a small proportion of the observations are excluded on a nonrandom basis, the bias will be small and safely ignored. There is to date no way of knowing in either case how small is small, in part because the importance of bias depends on the accuracy one needs for the particular analysis at hand. Perhaps the best advice is always to begin with the assumption that sample selection bias exists and proceed where possible with the corrections unless a strong argument can be made that moots the problem. For example, random assignment to treatment and control groups will insure internal validity, as long as there is no postassignment attrition, because the treatment dummy variables will, on the average, be orthogonal with the omitted sample selection variable (e.g., the hazard rate).

Corrections for sample selection bias often must overcome many practical difficulties. For example, while the set of regressors for the substantive and selection equations may within the Heckman framework be identical, the result (in these circumstances) will always be

[13] When a regressor is highly collinear with the intercept, only the standard error of the intercept is inflated (see Belsley et al., 1980, for a state-of-the-art discussion of multicollinearity).

[14] The correlation cannot be directly estimated in a consistent manner (Heckman, 1976).

very high multicollinearity in the substantive equation between the hazard rate and the other regressors. Indeed, were it not for the nonlinear probit form, the regression parameters in the substantive equation would be underidentified.

Finally, it must be emphasized again that this paper is no more than an introduction to the issues. There is a rich technical literature whose key articles should be read and a host of extensions with broad applicability: sample selection bias under multiple selection processes (e.g., Klepper et al., 1981), sample selection bias in simultaneous equation models (e.g., Amemiya, 1974; Sickles and Schmidt, 1978), sample selection bias in experimental designs (e.g., Barnow et al., 1980), sample selection bias in confirmatory factor analysis (Muthen and Jöreskog, 1981), and others. Many have immediate implications for sociological research.

REFERENCES

Amemiya, Takeshi
1974 "Multivariate regression and simultaneous equation models when the dependent variables are truncated normal." Econometrica 42:999–1011.
1981 "Qualitative response models: a survey." Journal of Economic Literature 19:1483–1536.

Arabmazar, Abbas and Peter Schmidt
1982 "An investigation of the robustness of the Tobit estimator to non-normality." Econometrica 50:1055–68.

Barnow, Burt S., Glend G. Cain and Arthur S. Goldberger
1980 "Issues in the analysis of selectivity bias." Pp. 43–59 in E. Stromsdorfer and G. Farkus (eds), Evaluation Studies Review Annual, Vol. 5. Beverly Hills: Sage.

Belsley, David A., Edwin Kuh and Roy E. Welsh
1980 Regression Diagnostics. New York: Wiley.

Berk, Richard A., Sarah F. Berk, Donileen R. Loseke and David Rauma
1983 "Mutual combat and other family violence myths." In David Finkelher, Richard J. Gelles, Gerald T. Hotaling, Murray A. Straus (eds.), The Dark Side of Families. Beverly Hills: Sage.
1981 "Throwing the cops back out: the decline of a program to make the criminal justice system more responsive to incidents of family violence." Social Science Research 11:245–79.

Berk, Richard A. and Subhash C. Ray
1982 "Selection biases in sociological data." Social Science Research 11:301–40.

Berk, Richard A. and Anthony Shih
1982 "Measuring citizen evaluations of the criminal justice system: a final report to the National Institute of Justice." Santa Barbara, CA: SPRI.

Berk, Richard A., Peter H. Rossi and Kenneth J. Lenihan
1980 "Crime and poverty: some experimental evidence for ex-offenders." American Sociological Review 45:766–800.

Dillman, Donald A.
1978 Mail and Telephone Surveys: The Total Design Method. New York: Wiley.

Engle, Robert F., David F. Hendry and Jean-Francois Richard
1983 "Exogeneity." Econometrica 51:277–304.

Goldberger, Arthur S.
1981 "Linear regression after selection." Journal of Econometrics 15:357–66.

Greene, William H.
1981 "On the asymptotic bias of the ordinary least squares estimator of the Tobit model." Econometrica 49:505–13.

Heckman, James J.
1976 "The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models." Annals of Economic and Social Measurement 5:475–92.
1979 "Sample selection bias as a specification error." Econometrica 45:153–61.
1980 "Sample selection bias as a specification error." Pp. 206–48 in James P. Smith (ed.), Female Labor Supply: Theory and Estimation. Princeton: Princeton University Press.

Judge, George G., William E. Griffiths, Carter R. Hill and Tsoung-Chao Lee
1980 The Theory and Practice of Econometrics. New York: Wiley.

Klepper, Steven, Daniel Nagin and Luke Tierney
1981 "Discrimination in the criminal justice system: a critical appraisal of the literature and suggestions for future research." Social Science Department, Carnegie-Mellon University.

Kmenta, Jan
1971 Elements of Econometrics. New York: MacMillan.

Lawless, Jerald F.
1982 Statistical Models and Methods for Lifetime Data. New York: Wiley.

Lord, F. M. and M. R. Novick
1968 Statistical Theories of Mental Test Scores. Reading, PA: Addison-Wesley.

Muthen, Bengt and Karl G. Jöreskog
1981 "Selectivity problems in quasi-experimental studies." Mimeo, Department of Statistics, University of Uppsala.

Olsen, Randall
1980a "Approximating a truncated normal regression with the method of moments." Econometrica 48:1099–1105.
1980b "A least squares correction for selectivity bias." Econometrica 48:1815–20.

Pearson, Karl and Alice Lee
1908 "Generalized probable error in multiple normal correlation." Biometrica 6:59–68.

Pindyck, Robert S. and Daniel L. Rubinfeld
1981 Econometric Models and Economic Forecasts. Second Edition. New York: McGraw Hill.

Ray, Subhash C., Richard A. Berk and William T. Bielby
1980 "Correcting for sample selection bias for a bivariate logistic distribution of disturbances." Paper presented at the 1980 meetings of the American Statistical Association.
Rossi, Peter H., Richard A. Berk and Kenneth J. Lenihan
1980 Money, Work, and Crime: Some Experimental Results. New York: Academic Press.
Rossi, Peter H., James D. Wright and Eleanor Weber-Burdin
1982 Natural Hazards and Public Choice. New York: Academic Press.
Sickles, Robin C. and Peter Schmidt
1978 "Simultaneous equation models with truncated dependent variables: a simultaneous Tobit model." Journal of Economics and Business 33:11–21.
Tobin, James
1958 "Estimation of relationships for limited dependent variables." Econometrica 26:24–36.
Tuma, Nancy B.
1982 "Nonparametric and partially parametric approaches to event-history analysis." Pp. 1–60 in Samuel Leinhardt (ed.), Sociological Methodology, 1982. San Francisco: Jossey-Bass.
Tuma, Nancy B., Michael T. Hannan and Lyle P. Groenveld
1979 "Dynamic analysis of event histories." American Journal of Sociology 84:820–54.

ORDINAL MEASURES IN MULTIPLE INDICATOR MODELS: A SIMULATION STUDY OF CATEGORIZATION ERROR*

DAVID RICHARD JOHNSON JAMES C. CREECH
University of Nebraska

Categorization error occurs when continuous variables are measured by indicators with only a few categories. When several continuous variables are collapsed into ordinal categories, the measurement errors introduced may be correlated. This condition violates the assumptions of classical measurement theory. Using simulated data and a multiple indicator approach, we examine the problems that surround categorization error. In general, our results indicate that while categorization error does produce distortions in multiple indicator models, under most of the conditions that we explored, the bias was not sufficient to alter substantive interpretations and the estimates were efficient. Caution is warranted in the use of two-, three- or four-category ordinal indicators, particularly when the sample size is small, as the estimates tend to be biased and inefficient.

There has been an increasing trend for researchers to treat ordinal data as if they were measured at the interval level, using statistical techniques which assume interval measures. A number of methodological studies have investigated the consequences of this practice, with mixed results (see Bollen and Barb, 1981, and Henry, 1982 for good summaries of these results). A frequent strategy of these studies is to focus on the relationship between an underlying continuous variable and its rank-order counterpart (Labovitz, 1967, 1970; O'Brien, 1981a, 1982); the epistemic correlation between the two indicates how well the categorized version serves as a stand-in for the unmeasured continuous variable. More recently, attention has shifted to how ordinal data affect bivariate and multivariate analyses (Bollen and Barb, 1981; Kim, 1975, 1978) by comparing the relationship between categorized variables with the relationship between their continuous analogs.

Since so much of sociological data is based on forced-choice, Likert-type responses, it is surprising that the possibility of correlated measurement error among variables that have been collapsed, particularly when they have been collapsed in the same manner, has not been explored. Measurement theory is built on

*Direct all correspondence to: David R. Johnson and James C. Creech, Department of Sociology, University of Nebraska, Lincoln, NE 68588.

We are grateful to Kenneth Bollen, Hugh P. Whitt and Moshe Semyonov and two anonymous reviewers for comments made on previous drafts of this paper. This research was supported, in part, by a National Institute of Aging predoctoral traineeship provided by the Midwest Council for Social Research in Aging, granted to the second author.