DOCUMENT RESUME

ED 202 979 UD 021 501

AUTHOR Barnow, Burt S.; And Others
TITLE Issues in the Analysis of Selectivity Bias. Discussion Papers. Revised.
INSTITUTION Wisconsin Univ., Madison. Inst. for Research on Poverty.
SPONS AGENCY Department of Health, Education, and Welfare, Washington, D.C.; National Science Foundation, Washington, D.C.
REPORT NO IRP-DP-600-80
PUB DATE Apr 80
GRANT SOC-76-24428
NOTE 34p.; Some tables may be marginally legible due to reproduction quality of original document.

EDRS PRICE MF01/PC02/Plus Postage.
DESCRIPTORS Control Groups; *Evaluation Methods; Experimental Groups; *Models; *Program Evaluation; Research Methodology; *Statistical Bias

ABSTRACT

Selectivity bias arises in program evaluation when the treatment or control status of the subjects is related to unmeasured characteristics that themselves are related to the program outcome under study. This situation has the potential to lead to an incorrect estimation of the treatment effect when assignment to treatment and control groups is not random. This paper adopts techniques that were recently developed in the econometric analysis of labor and markets and applies these techniques in order to resolve the bias problem of conventional evaluation models. The selection bias problem is placed in a broader context by considering references to the institutional setting of evaluation research and to its application in labor economies. The paper concludes with several caveats about the proposed technique. (Author/MK)

University of Wisconsin-Madison

Institute for Research on Poverty

Discussion Papers

Burt S. Barnow
Glen G. Cain
Arthur S. Goldberger

ISSUES IN THE ANALYSIS OF SELECTIVITY BIAS

DP #600-80

Issues in the Analysis of

Selectivity Bias

Burt S. Barnow
Acting Director, Office of Research on Development,
Employment and Training Administration,
U.S. Department of Labor

Glen G. Cain
Professor of Economics
University of Wisconsin-Madison

Arthur S. Goldberger
Professor of Economics
University of Wisconsin-Madison

April 1980

This is a slightly revised version of the paper presented at the August 1978 meeting of the Econometric Society, in Chicago. The research was supported in part by National Science Foundation Grant SOC 76-24428 and by funds granted to the Institute for Research on Poverty at the University of Wisconsin-Madison granted by the Department of Health, Education, and Welfare pursuant to the provisions of the Economic Opportunity Act of 1964.

ABSTRACT

Selectivity bias arises in program evaluation when the treatment (or control) status of the subjects is related to unmeasured characteristics that themselves are related to the program outcome under study. This situation potentially leads to a misestimation of the treatment effect. In this paper we adopt techniques recently developed in the econometric analysis of labor markets to the bias problem in a conventional evaluation model. A resolution of the problem emerges under assumptions that are reasonably general. References to the institutional setting of evaluation research and to applications in labor economies place the "selection bias" problem in a broader context. The paper concludes with several caveats about the proposed technique.

Issues in the Analysis of Selectivity Bias

1. INTRODUCTION

Selectivity bias arises in program evaluations when the treatment (or control) status of the subjects is related to unmeasured characteristics that themselves are related to the program outcome under study. The term "bias" refers to the potential misestimation of the effect of the treatment (or program) on the outcome. Selectivity bias is a concern whenever the assignment to treatment and control groups is not random, conditional on whatever observable explanatory variables, if any, are used in the analysis. So stated, it is clear that the issue of selection bias is pervasive in empirical research in economics because the assignment of observations to the different statuses defined by the predictor or explanatory variables of interest are seldom explicitly random. Thus, there should be a common ground in the methods used to analyze selectivity bias in program evaluation and econometrics, distinct but increasingly overlapping fields.

In this paper we adapt techniques developed in the econometric analysis of labor markets--particularly by James J. Heckman (1974, 1976, 1978, 1979), G. S. Maddala (1976, 1978) and Maddala and Lung-fei Lee (1976)--to the bias problem in a conventional evaluation model. A resolution of the problem emerges under assumptions that are reasonably general.

The next section of the paper provides an institutional background by describing developments in program evaluation and in applied labor

economics that deal with and illustrate the issue of selection bias. Section 2 then introduces a simple model that illustrates formally the statistical issues, notes conditions when no bias exists and when the direction of bias is known, and points to several misunderstandings in the program evaluation literature. Section 3 surveys several econometric approaches in labor economics that have attempted to eliminate (or, equivalently, to quantify) the bias. Section 4, which is the core of the paper, applies the new econometric approach to the conventional evaluation model to indicate how the selection bias in this model may be resolved.

1. Developments in Program Evaluation and Labor Economics Involving Selection Bias

Evaluation research began with the analysis of rather small-scale projects, mainly in medicine, experimental and social psychology, educational psychology, and economics. In economics, the basic approach was benefit-cost analysis, which was different from the more direct application of statistics in the other disciplines. Since the mid-1960s evaluation research has been applied to large-scale governmental programs whose main purpose was to improve the well-being of a sizeable segment of society. Evaluation was not the main objective of the programs, which largely explains why random assignment to treatment and control groups was almost never employed. Nevertheless, ex post evaluations were publicly demanded because the successes or failures of the programs were major political issues.

The current practice of program evaluation research by economists is quite different from the ex ante evaluation of water resource projects and the like, which were the mainstay of traditional benefit-cost analysis.[1]

In ex post evaluations, the first priority is to estimate the quantitative effect of the program, which logically precedes the determination of whether the benefits exceed the costs. This priority puts a greater emphasis on statistical models and has brought economic evaluation research closer to the approaches used in other disciplines.

Today, evaluation research is a huge enterprise in applied social science. The Evaluation Research Society was organized in 1976, and the journal Evaluation began in 1973 with a grant from the National Institute of Mental Health. Courses on the topic are offered in graduate schools of many universities, often in recently established departments or centers of public administration and policy research. A two-volume Handbook of Evaluation Research, intended as a text for graduate courses, includes a bibliography of some 1500 items, almost all post-1960 and almost all by noneconomists.[2] Other books on evaluation, filling several shelves in one's bookcase, also tend to be recently published and written by noneconomists.[3]

The volume Federal Program Evaluation contains an inventory of approximately 1700 evaluation reports produced by and for 18 selected Federal agencies, covering the period 1973-1975. Many were written by researchers outside the government, but all were related to agency programs.[4] One may safely credit (or blame) the federal government, beginning with the Great Society programs of the Johnson administration, for making evaluation research the growth industry that it now is.

The topic of our paper, selectivity bias, occupies a small part of the statistical methodology aspect of evaluation research, an aspect which is, itself, only a small part of the total field. Nevertheless, selectivity

bias can be given a broad interpretation, with implications and applications extending beyond evaluation research. For example, selectivity bias can be viewed broadly as a version of specification error in statistical models in which behavioral outcomes are functions of "predictor" or "explanatory" variables. In this version, the outcome of a program is examined in the same way as an outcome of a controlled experiment in a laboratory or, at the other end of the spectrum of research settings, as an outcome in a historical process measured with time-series data. There is a common framework used to measure the causal effect on the outcome of, respectively, the program, the laboratory treatment, or the historical event.

Selectivity concerns the presence of some characteristic of the treatment (or control) group that is associated both with receipt of the treatment and also with the outcome so as to lead to a false attribution of causality regarding treatment and outcome. So stated, selectivity bias is a version of omitted-variable bias, commonly analyzed under the rubric of specification error in econometric models. The unbiased measurement of causal effects is a broad and complex topic. As the eminent statistician W. Edwards Deming (1975) stated, "Evaluation is a study of causes."

In labor economics the issue of selection bias in evaluation models has been confronted directly in a number of recent studies. Training and education programs deal with individuals who are either selected for the program or who are self-selected. If this selection process is not fully known to the investigator, an unbiased measure of the treatment may be unobtainable. For explicit attention to this problem, see the evaluation study of a government training program by Orley Ashenfelter

(1973) and the evaluations of compensatory education programs by Irwin Garfinkel and Edward M. Gramlich (1973) and by Burt S. Barnow and Glen G. Cain (1977). Regulatory programs usually deal with firms, and again there is selection by the agencies (or self-selection) for participation in the regulatory process. The apparent effects of the program may be biased because the outcomes may reflect unmeasured preexisting characteristics of the selected firms. Evaluations of anti-discrimination and affirmative action programs are a case in point, and Heckman and Kenneth I. Wolpin's (1976) study of the Office of Federal Contract Compliance of the U.S. Department of Labor deals explicitly with selection bias. Another example is the study by Robert S. Smith (1975) of the impact on injuries of the "target industries program" of the Occupational Safety and Health Act (see also the critique by Jack E. Triplett [1975]).

There are also well-known examples of the selection bias issue in other areas of labor economics, testifying to the pervasiveness of the issue in empirical economic research. Attempts to measure the effect of unions on wages are especially interesting because the selection process is so varied in several dimensions. The units of observation may be industries, occupations, firms, or individuals. There is an element of self-selection among individual workers, dependent in part on the worker's preferences for unionism; there is selection by union organizers, dependent in part on the costs and benefits to the union of organizing the work place; and there is selection by employers through their hiring and personnel policies. Some twenty years ago H. Gregg Lewis (1959) discussed these potential selection biases in examining the union wage effect. Several

recent econometric studies have attempted to model the selection of workers into union status in the course of estimating the union wage effect: Ashenfelter and George Johnson (1972), Peter Schmidt and Robert P. Strauss (1976), Lee (1978), and Duane Leigh (1978) (see also Randall J. Olsen [1978]). The intricacies in this area are illustrated by noting that the bias may differ across demographic or social groups in the population. Ashenfelter (1972) estimated a larger union effect for blacks than for whites, and Zvi Griliches (1976) estimated a relatively large union effect on the wages of young men, ages 17-27. Does the selection process operate differently among such groups of workers, and, if so, are differential biases a consequence?

Labor supply studies of women have probably provided the most explicit attention to the selectivity bias issue in economics. Here the market wage of women is the analogue to the treatment in evaluation studies, and the selectivity process affects labor force participation and thus the observability of the market wage. A special feature here is that the sample for which a wage is measured is truncated--no wage is measured for non-working women. In the typical program-evaluation design, the treatment variable is measured for both the treatment and control groups. Heckman provided an early analysis of this problem (1974) and a useful review article (1976). Yoram Ben-Porath (1973), Reuben Gronau (1973, 1974), and Lewis (1974) had earlier discussions of this application of selectivity bias, and T. Paul Schultz (1980) and Giora Hanoch (1980) have contributed more recent work.

Even those rare examples of economic research which use controlled experiments with random assignments to program status--namely, the negative income tax experiments--have been forced to confront selectivity bias. The issue arises with respect to attrition, particularly when the subjects leave the experiment for one of the existing welfare programs. Analyses of these experiments have included attempts at modeling the processes of attrition and of participation in welfare programs: Harold W. Watts, Jon K. Peck, and Michael Taussig (1977), and Garfinkel (1977).

Other studies in labor economics will be cited in Section 3 as illustrative of specific approaches to dealing with selectivity bias. Our thesis that the issue of selectivity bias is important and pervasive in econometric research as well as in program evaluation research should be noncontroversial.

2. A SIMPLE FORMULATION OF THE EVALUATION MODEL

We focus on that part of an evaluation that seeks to measure the effect of the program on a specific quantified outcome. Many aspects of a full evaluation are ignored in this narrow focus: the costs of the program, the dollar-equivalent value of the outcome, the administration of the program, the equity issues involved in the distribution of benefits and costs, the correspondence between measured outcomes and political objectives, the question of multiple objectives, and others.

For simplicity's sake, assume that the outcome, y, is linearly related to the treatment status, z (defined by participation in the program), and to an unobserved variable, w, defined as the preprogram "true ability" to achieve the outcome. A pure random term, ε, completes the equation. In

this hypothetical evaluation model, with y systematically related to z and w, there would be no need for any other control variables. The evaluator's interest is in the effect of z on y, given true ability, and by assumption w completely measures that true ability. In (1) below, α is the true treatment effect:

$$(1) \qquad y = \alpha z + w + \varepsilon.$$

But this equation is nonoperational because w is unobserved (which, incidentally, is why assigning it a unit coefficient is innocent). How may the evaluator persuade an interested audience that the measured effect of z on y is free of any contamination from a correlation between z and w, given that w is not available as an explanatory variable? Random assignment to the z status is convincing in principle. But the integrity of randomization may be compromised in practice (by reliance on volunteers, by the absence of a double-blind design, by attrition), and in any event, almost all programs deliberately use nonrandom assignments.

There should be widespread agreement among economists that random assignments are not essential to the estimation of unbiased treatment effects.[5] As we have argued in earlier papers, unbiasedness is attainable when the variables which determined the assignment are known, quantified, and included in the equation: Goldberger (1972), Barnow (1973), Cain (1975). Assume that an observed variable, t, was used to determine assignment into the treatment group (z = 1) and the control group (z = 0). In general, t, which we refer to as the selection variable, would be a score based on a composite of variables, some of which would be correlates of

ability, w. Nevertheless, since t is the only systematic determinant of treatment status, t will capture any correlation between z and w. Thus, the observed t could replace the unobserved w as the explanatory variable in (1). In equation (2), β_1 would be unbiased, that is, equal to x:

$$(2) \qquad y = \beta_1 z + \beta_2 t + \varepsilon^*$$

The use of either w or t as an explanatory variable, then, will free z from the contamination which leads to selectivity bias.

"Modeling the selection process" is, of course, precisely what one claims to do when specifying a multiple regression which holds constant those traits of the unit of observation that affect the outcome and that are correlated with the input variable of interest. A purely random determinant of assignment would be harmless, where "random" refers to a selection variable--such as the flip of a coin--that does not affect the outcome. Theory is supposed to tell us which selection variables are associated with outcomes. As examples: how do persons get "selected into" different educational attainment categories or into union membership, and how do the variables fully describing this process relate to an outcome variable like earnings?

To illustrate briefly the selection model with an example that has spawned an extended controversy, assume z is participation in the Head Start compensatory education program, y is the postprogram test score that is presumed to measure cognitive echievement, and t represents the family income of the children. Assume, further, that those children for whom t is below the poverty line (t_p) are in the program (z = 1) and those with values of t above t_p are excluded (z = 0). Even though the correlation

between w and t and the correlation between z and t are both negative, there is no bias in $\hat{\beta}_1$ in model (2). Figure 1 illustrates the magnitude of $\hat{\beta}_1$ in the case that the program is beneficial.

A misunderstanding has arisen in the uses and interpretation of this model. In a seminal paper on evaluation methodology, the psychologist Donald T. Campbell (1969) suggested that this model, which he calls the "regression discontinuity" design, is severely restricted. But his reasons, in fact, do not apply. Campbell and Albert Erlebacher (1970) say that the model requires a random assignment among "ties" on the boundary line of participation (t_p in our example), and that "we would learn about the effects of the program only for a narrow band . . . . We would wonder about its effectiveness for the most disadvantaged."[6] But model (2) uses the full range of t; ties are inconsequential, and therefore no randomization is needed; and the entire range of values of t provides potential information for the effectiveness of the program, even for nonlinear or interactive effects. Figures 2A and 2B illustrate negative and positive interactions between the treatment and t. The nonlinear functional forms shown are chosen to avoid both deleterious and explosive treatment effects when t becomes large; such nonlinearities, of course, would have to be specified in the model.

Let us examine next the case where the selection process is not known precisely, in the sense that the available data do not permit quantification of t. Assume that there is available a vector $\underline{x}$, composed of variables that are correlated with--that is, proxy for--ability, w. At the same time, $\underline{x}$ may include variables which enter t. The equation to be fitted is

Figure 1

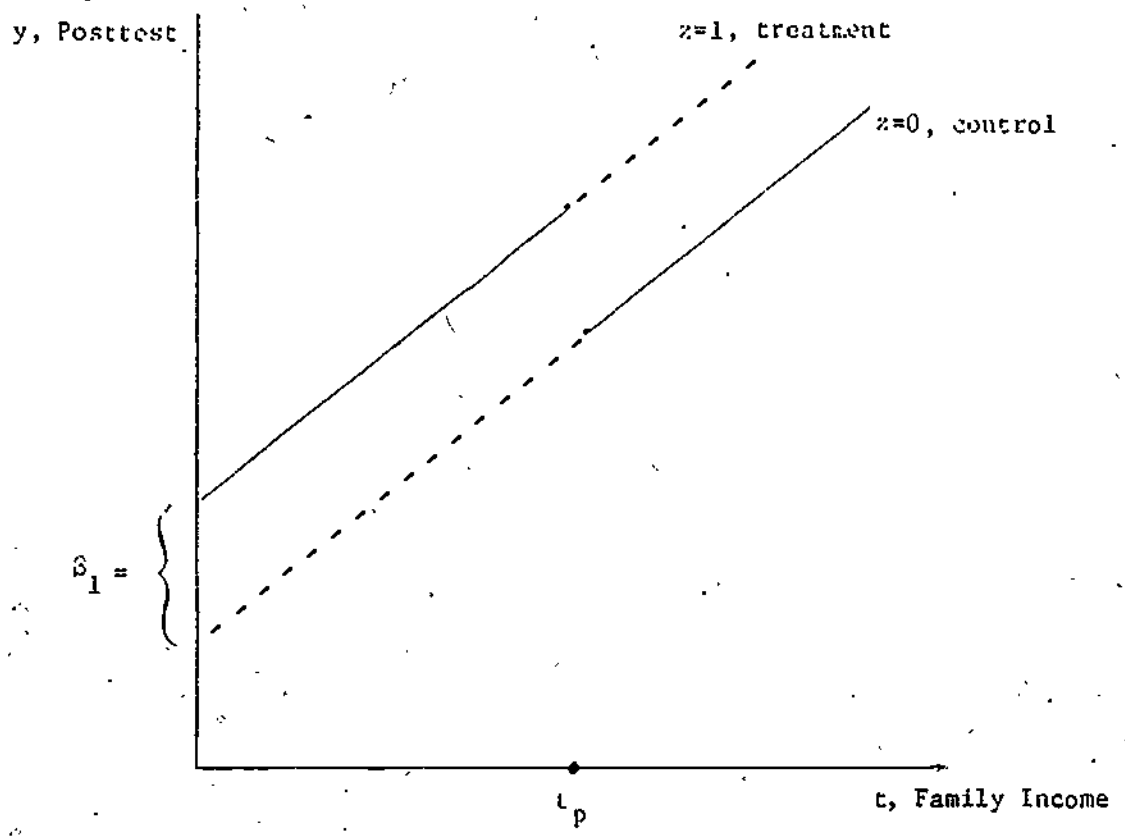


Figure 1. Nonrandom assignment to treatment and control groups, based on family income, t. Dashed lines represent nonobserved extrapolation of y, given t and the treatment/control status, z.

Figure 2A

Negative z × t interaction

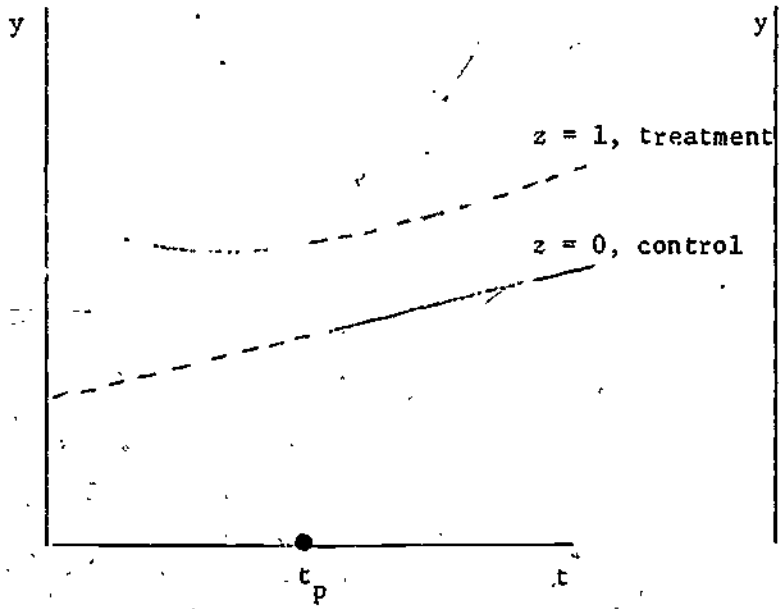
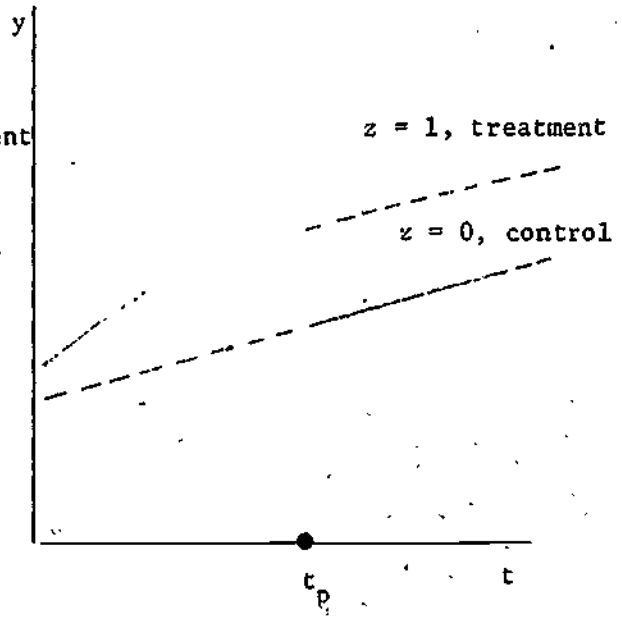


Figure 2B

Positive z × t interaction



Figures 2A and 2B — Nonrandom assignments to treatment and control groups based on family income, t, in examples where the treatment interacts with t. Dashed lines represent nonobserved extrapolations of y, given t and the treatment/control status. In Figure 2A the upper solid line, showing a negative interaction, is assumed to reach an asymptote rather than to decline indefinitely, thus avoiding a deleterious treatment effect at higher t values. Similarly in Figure 2B, a positive treatment x selection interaction increases at a decreasing rate and reaches an asymptote, thus preventing the treatment effect from exploding at higher t values.

(3) $$y = \gamma_1 z + \underline{\gamma}_2' \underline{x} + \varepsilon^{**}.$$

An estimate of γ_1 in this model will in general be biased for α. As discussed in our earlier papers, the bias depends on the covariance of z and w conditional on $\underline{x}$ and may be positive or negative. On this point a misunderstanding arose in the well-known and often-cited paper by Campbell and Erlebacher (1970) dealing with an evaluation of Head Start. They argued that the direction of bias could be inferred on the basis of the $\underline{x}$, z correlation:

> How can one tell which direction a matching bias will take? Only by having evidence of the nature of the population difference which matching attempted to overcome . . . . This undermatching [on true ability] showed up on the socioeconomic status ratings subsequently made [which showed a negative correlation between socioeconomic status and treatment status].[7]

It is the last sentence in the quotation that gave rise to the misunderstanding. In the example discussed, $\underline{x}$ was the socioeconomic status of the parents. Knowing that the $\underline{x}$, w correlation is positive, one cannot infer that a negative $\underline{x}$, z correlation biases down the Head Start (or z) effect. What determines the bias is the <u>conditional</u> covariance, $C(z,w|\underline{x})$ and not the unconditional covariance, $C(z,w)$.[8]

Consider the general empirical model represented by equation (3) above, in which the investigator has measures of $\underline{x}$. As one polar case, $\underline{x}$ coincides with w, and there would be no bias in estimating the treatment effect regardless of the selection process. Here, $C(z,w) \neq 0$ but $C(z,w|\underline{x}) = 0$ because $\underline{x}$ coincides with w. This polar case is presumably hypothetical because w is unobserved. As another polar case, $\underline{x}$ coincides with t, and

there is again no bias, as was discussed above in connection with equation (2). Here again $C(z,w) \neq 0$, but $C(z,w|\underline{x}) = 0$ because selection on t (= $\underline{x}$) means that the only relation between w and z is that which is induced by the relation between w and t (= $\underline{x}$), a source of correlation which is fully captured by $\underline{x}$. A third polar case is that t, and therefore the z status, is random, like a coin flip. Here again, $C(z,w|\underline{x}) = 0$, and indeed $C(z,w) = 0$.

This leaves the important case when w is unobserved and the selection is nonrandom and not fully measured. With no other outside information, nothing can be said about the direction of bias. With information about the conditional covariances--in particular $C(z,w|\underline{x})$, but also $C(t,\underline{x}|w)$ or its equivalent in sign, $C(z,\underline{x}|w)$--direction of bias may be determined. With information about the distribution of the unobserved variables, t and w, along with information about the functional form relating t and w to $\underline{x}$, it turns out--perhaps surprisingly--that the bias may be quantified, and further, that this achievement does not require information about the signs of the relations, conditional or unconditional, among t, w, z, and $\underline{x}$, beyond what may be directly measured on the basis of the observed values of z and $\underline{x}$. The quantification (or equivalently the elimination) of bias in terms of the model of evaluation is demonstrated in Section 4. Let us first survey briefly the handling of selectivity bias in the recent econometric studies in labor economics.

3. APPROACHES TO SELECTION BIAS IN LABOR ECONOMETRICS

As we read the labor econometrics literature, it taps additional information about the selection process to obtain an unbiased estimate of the

treatment effect when the observed $\underline{x}$-variables exhaust neither t nor w. In essence, a selection equation with z as the dependent variable is specified, and restrictions are placed on it relative to the outcome equation--our (3). The role of the restrictions is to purge the apparent treatment effect--γ_1 in our equation (3)--of any preexisting differences between the treatment and control groups. In our reading, we detect two types of restrictions. The first type specifies that one or more variables determining selection do not affect the outcome and hence are excludable from the outcome equation. Thus, information on variables excluded from $\underline{x}$ is required. The second type specifies a functional form for the relation between $\underline{x}$ and w and a nonlinear relation between z and $\underline{x}$. This leads to a nonlinear function of $\underline{x}$ in the outcome equation, which serves to control for any z, w relation that is net of $\underline{x}$.

In distinguishing these two types of restrictions, we pass over what is no doubt the most common approach: simply assume away the selection bias after a diligent attempt to include a large number of variables in $\underline{x}$ that control for ability, w. The argument, or assumption, is that whatever the selection variables may be, beyond those included in $\underline{x}$, they are unrelated to outcome. In the two-equation recursive formulation for z and for y, which we discussed above and present formally in section 4, this amounts to assuming that z is determined by both $\underline{x}$ and a disturbance that is uncorrelated with the disturbance, ϵ^{**}, in equation (3) determining y. Most of the research estimating the effect of education (the treatment variable) on earnings (the outcome) adopts this approach. Also, the empirical estimates of the Head Start effect on test scores by Barnow and Cain (1977) fell in this category.

Two examples suffice to illustrate the first type of restriction, which is to obtain additional $\underline{x}$'s that have no effect on the outcome and then to use this information to identify the treatment effect. A trivial but universally accepted case is the coin flip. When the t variable is generated naturally by environmental or market forces, however, the assumption that the "selection variables" have no effect on the outcome will usually be controversial. For example, in a study of the effect of unionism on wage rates, Ashenfelter and Johnson (1972) variously assumed that the concentration ratio of the industry or the region of residence (South or non-South) determines the extent of unionism, while concentration and region have no causal effect upon wage rates. Thus, those two variables are excluded from the structural equation for wages. The authors suggested that those identifying restrictions might be considered arbitrary.[9]

The second type of restriction permits distinguishing between the way the $\underline{x}$-variables affect selection from the way they affect outcomes (or, equivalently, the way they represent ability). David Greenberg and Marvin Kosters (1973) analyzed the relation between labor supply (an outcome) and nonlabor income (the treatment). Their initial empirical work led them to specify a selection process that involved differential tastes or preferences for asset accumulation (and, thus, for nonlabor income). They used some of the $\underline{x}$-variables from the labor supply equation

to estimate asset preferences by the device of a regression of observed assets on the selected $\underline{x}$-variables. Predicted assets were included in the final labor supply equation, and identification of its effect was achieved by virtue of its nonlinear relation to $\underline{x}$. The controversial feature of this procedure is that the specification of the nonlinear function between observed assets and the set of $\underline{x}$-variables was somewhat ad hoc.

More recent work, beginning with Heckman (1974), appears to be considerably more general. A nonlinear functional form in $\underline{x}$ that serves to identify the treatment effect is not imposed directly but rather emerges from two assumptions: one, that the distribution of the error terms in the equations for the omitted variables, w and t, is bivariate normal; and, two, that the functional forms relating w and t to $\underline{x}$ are known--in practice, known to be linear. We develop this in Section 4.

Heckman's model, dealing with the labor supply of women, is complicated by two features: one, simultaneity, which does not appear in the typical evaluation model; and the other, truncation, which does appear but in a somewhat disguised form. First, in the women's labor supply model, the analogue to the omitted ability variable is the "shadow price of time" in housework, and this is assumed to be affected by the outcome variable, hours of market work. Thus, a simultaneous-equation model is specified and, as with the Ashenfelter and Johnson model, additional identifying restrictions are required. Second, as mentioned earlier, the analogue to the treatment variable is the wage rate, and this is not observed for women who supply zero hours of market work. Thus, the distribution of wage rates is truncated and, unlike the treatment variable in program evaluation models, is not

measured for all observations. This is only an apparent difference, however. In the evaluation model with differential ability between treatment and control groups, the two groups are effectively truncated (or censored) on ability, and the technique for correcting for the omitted variable bias is the same as the technique for correcting for the truncation bias.

4. UNBIASED ESTIMATION OF TREATMENT EFFECTS IN THE EVALUATION MODEL

We now sharpen the specification of the evaluation model in a way which will permit unbiased estimation of the treatment effect. The observable variables are y = outcome, z = treatment (z = 1 for treatment group, z = 0 for control group), and $\underline{x}$ = vector of covariates (including the constant). The unobserved variables are w = ability and t = selection, along with various disturbances to be introduced below. By definition of t, assignment to the two groups is determined by

$$z = \begin{cases} 1 & \text{if } t > 0 \\ 0 & \text{if } t \leq 0. \end{cases} \tag{4}$$

By definition of w, outcome is determined by

$$y = w + \alpha z + \varepsilon_o , \tag{5}$$

where the disturbance ε_o is normally distributed, independent of w and z, with expectation zero and variance σ_{oo}.

Consider the joint probability distribution of w, t, and $\underline{x}$ in the initial population, that is, prior to selection and treatment. We suppose that

$$w = \underline{\theta}_1' \underline{x} + \varepsilon_1 \tag{6}$$

(7) $$t = \underline{\theta}_2' \underline{x} + \varepsilon_2$$

where $\underline{\theta}_1$ and $\underline{\theta}_2$ are coefficient vectors and where the disturbances ε_1 and ε_2 are bivariate-normal, independent of $\underline{x}$ (and of ε_o) with expectations zero, variances σ_{11} and σ_{22}, and covariance σ_{12}. Note that in (6)-(7) two sources of nonrandom selection are distinguished: w and t can be correlated via their common dependence on $\underline{x}$ and via the correlation of their disturbances ε_1 and ε_2. Equations (6)-(7) are not intended to be causal: we need not say that $\underline{x}$ determines w and/or t. Rather, the specification may be interpreted as purely descriptive of the joint probability distribution of w, t, and $\underline{x}$.

Proceeding to the analysis, we substitute (6) into (5) to get the outcome equation:

(8) $$y = \underline{\theta}_1' \underline{x} + \alpha z + \varepsilon_3 ,$$

where $\varepsilon_3 = \varepsilon_1 + \varepsilon_o$. Now ε_2 and ε_3 are bivariate-normal, independent of $\underline{x}$, with expectations zero, variances σ_{22} and $\sigma_{33} = \sigma_{11} + \sigma_{oo}$, and covariance $\sigma_{23} = \sigma_{12}$. We seek $E(y|\underline{x}, z)$, the expectation of outcome conditional on the covariates and the treatment dummy.

From (4) and (7) we see that the event $z = 1$ is equivalent to the event $\underline{\theta}_2' \underline{x} + \varepsilon_2 > 0$, and thus to $\varepsilon_2 > - \underline{\theta}_2' \underline{x}$, and thus to $(\varepsilon_2/\sigma_2) > - \underline{\theta}'\underline{x}$, where $\sigma_2 = \sqrt{\sigma_{22}}$ and $\underline{\theta} = (1/\sigma_2)\, \underline{\theta}_2$. By the same argument, the event $z = 0$ is equivalent to $(\varepsilon_2/\sigma_2) \leq - \underline{\theta}'\underline{x}$. Now ε_2/σ_2 is a standard normal variable independent of $\underline{x}$. Since z is binary, it follows that

(9) $$E(z|\underline{x}) = \text{Prob}\{z = 1|\underline{x}\} = 1 - F(- \underline{\theta}'x) = F(\underline{\theta}'\underline{x}),$$

where $F(\cdot)$ denotes the standard normal cumulative distribution function. Further, it follows that

$$E((\varepsilon_2/\sigma_2)|\underline{x},\ z = 1) = f(\underline{\theta}'\underline{x})/F(\underline{\theta}'\underline{x}), \tag{10}$$

$$E((\varepsilon_2/\sigma_2)|\underline{x},\ z = 0) = -f(\underline{\theta}'\underline{x})/(1 - F(\underline{\theta}'\underline{x})), \tag{11}$$

where $f(\cdot)$ denotes the standard normal density function: see Johnson and Kotz (1970, p. 81). Using f and F as shorthand for $f(\underline{\theta}'\underline{x})$ and $F(\underline{\theta}'\underline{x})$ respectively, we can assemble (10)-(11) into, say,

$$\begin{aligned} E((\varepsilon_2/\sigma_2)|\underline{x},\ z) &= z\,f/F + (1 - z)f/(1 - F) \\ &= \left(\frac{f}{F(1 - F)}\right)(z - F) \\ &= h(\underline{x},\ z;\ \underline{\theta}). \end{aligned} \tag{12}$$

Then $E(\varepsilon_2|\underline{x},\ z) = \sigma_2'\, h(\underline{x},\ z;\ \underline{\theta})$, and with the distributional information under (8), it follows that

$$E(\varepsilon_3|\underline{x},\ z) = (\sigma_{12}/\sigma_{22})\ E(\varepsilon_2|\underline{x},\ z) = \mu h(x,\ z;\ \vartheta), \tag{13}$$

where $\mu = \sigma_{12}/\sigma_2$: see Johnson and Kotz (1972, p. 112).

In view of (13), the expectation of (8) conditional on $\underline{x}$ and z is

$$E(y|\underline{x},\ z) = \underline{\theta}_1'\underline{x} + \alpha\ z + \mu\ h(\underline{x},\ z;\ \underline{\theta}). \tag{14}$$

Since this is a conditional expectation function relating observable variables, its parameters, namely $\underline{\theta}_1$, α, μ, and $\underline{\vartheta}$, are consistently estimable by nonlinear least squares. We have thus established a method of obtaining an unbiased,

or, to be precise, a consistent estimate of the treatment effect in our specification of the evaluation model. In doing so, we have simply restated the arguments in Heckman (1976) and Maddala and Lee (1976).

To estimate (14) in practice, a two-step procedure may be used. First, estimate $\underline{\theta}$ by maximum-likelihood probit analysis of z on $\underline{x}$, and insert those estimates $\hat{\underline{\theta}}$ in place of $\underline{\theta}$ in (12) to calculate $\hat{h} = h(\underline{x}, z; \hat{\underline{\theta}})$ at each observation. Second, estimate $\underline{\beta}_1$, α, and μ by linear least squares regression of y on $\underline{x}$, z, and $\hat{h}$. This too provides consistent estimates.

A main theme in the evaluation research literature on selection bias is thus verified: linear regression of y on $\underline{x}$ and z produces biased estimates of α, the treatment effect. But in the present formulation the precise source of the bias is apparent, namely omission of the $h(\underline{x}, z; \underline{\theta})$ variable in (14). Once this term is included, least-squares regression gives a proper estimate of α. We observe from (14) that linear regression of $\underline{x}$ and z alone would give an unbiased estimate of α in the special case $\mu = 0$, that is $\sigma_{12} = 0$, that is $C(w, \varepsilon|\underline{x}) = 0$ or, equivalently, $C(w, z|\underline{x}) = 0$. This verifies the results discussed informally in sections 2 concerning the absence of bias when assignment is purely random, or purely on the basis of the observable covariates.[10]

A still simpler approach is also available. With the aid of (9) and (12) we recognize that conditional on $\underline{x}$, $h(\underline{x}, z; \underline{\theta})$ is a constant times $z - E(z|\underline{x})$, so that

$$E[h(\underline{x}, z; \underline{\theta})|\underline{x}] = 0. \tag{15}$$

Applying the iterated expectation rule to (14) then gives

$$(16) \qquad E(y|\underline{x}) = \underset{z|x}{E}\,(E(y|\underline{x}, z)) = \underline{\theta}_1'\underline{x} + \alpha E(z|\underline{x}) = \underline{\theta}_1'\underline{x} + \alpha F(\underline{\theta}'\underline{x}).$$

Indeed, (16) can be obtained directly from (8): see Maddala and Lee (1976, p. 528) and thus holds without assuming normality for ε_0. Now (16) is also a conditional expectation function relating observable variables, so its parameters, $\underline{\theta}_1$, α, and $\underline{\theta}$ are consistently estimated by nonlinear least squares. In practice a two-step procedure may be used: First, estimate $\underline{\theta}$ by maximum-likelihood probit analysis of z on $\underline{x}$, and insert those estimates in place of θ to calculate $\hat{z} = F(\hat{\underline{\theta}}'\underline{x})$ at each observation. Second, estimate $\underline{\theta}_1$ and α by linear least squares regression of y on $\underline{x}$ and $\hat{z}$. From this perspective, linear regression of y on $\underline{x}$ and z produces biased estimates of α because the variable z has not been purged of its endogeneity.

For more formal discussion of these and alternative estimation procedures, see also Takeshi Amemiya (1978), Heckman (1978) and Lee (1979).

We believe that this straightforward application of the Heckman-Maddala-Lee approach resolves in principle the problem of selectivity bias as it arose in evaluation research. Having reached that point, we must indicate that a number of serious problems require attention among which are the following:

(i) Choice among alternative consistent estimation procedures.

(ii) High degree of collinearity in the second-step regressions.

(iii) Robustness of estimators to non-normality of disturbances: see Crawford (1979).

(iv) Misspecification of original model, in which case the nonlinear terms $h(\underline{x}, z; \vartheta)$ and $F(\underline{\theta}'\underline{x})$ may be proxying for omitted variables and/or nonlinearities.

(v) Multiple selection rules: see Waldman (1979).

This listing includes both conceptual problems--particularly (v), (iv), and possibly (iii)--and problems of implementation--mainly (i), (ii), and (iii). We are not able at this time to assess the frequency or seriousness of these problems, so the list will have to stand by itself as a rather stark agenda for future analysis.

NOTES

[1]Two early influential books in this tradition are Roland N. McKean, Efficiency in government through systems analysis (New York: John Wiley, 1958), and Otto Eckstein, Water resources development: The economics of project evaluation (Cambridge, Mass.: Harvard University Press, 1958).

[2]E. L. Struening, and Marcia Guttentag, eds., Handbook of evaluation research, 2 vols., (Beverly Hills, Calif., 1975, Sage Publications). These volumes will be cited as HER I and Her II.

[3]See the appendix.

[4]Federal program evaluation, a directory for the Congress, by the U.S. General Accounting Office (Washington, D.C.: GPO, 1976). No information is given on the cost of these evaluations. Perhaps a conservative estimate of the average cost of an evaluation study is $50,000. This implies that $85 million was spent in the 1973-75 period.

[5]However, consider the following statement by the economist Alice Rivlin (1971, pp. 111-112):

> A valid experiment requires that individuals be assigned to treatment or control groups by a random selection process. Chance must enter. A government official may find it far more difficult to explain to the public that he is allocating a scarce resource on the basis of chance than to defend some other selection criterion such as need or merit or "first come, first serve." But if such criteria are used, those not selected cannot validly be compared with the treatment group to establish the effectiveness of the treatment, because the two groups may differ in important ways.

[6]Campbell (1969) is reprinted in HER I, pp. 71-99. Campbell and Erlebacher (1970) is reprinted in HER I, pp. 597-617. The quotation in the text is on p. 615 of HER I.

[7]HER I, p. 606.

[8]Campbell and Erlebacher also referred to the fact that the analysts in the study being criticized had a more difficult time finding the "most disadvantaged" children among the control group (HER I, p. 607). This proves nothing, however. The most disadvantaged children may be more difficult to recruit generally for either treatment or control groups. Or the most disadvantaged control children who were found may have been lower on true ability than the treatment children of the same measured status.

[9]Actually, their model was more elaborate than our two-equation recursive representation as they allowed for full simultaneity between union (selection) status and wage rates (outcomes). Thus, they required additional identifying restrictions to estimate the effect of wage rates on union status and used the restrictions that an industry's average educational level and percentage of female employees affect wage rates but not the extent of unionism. Clearly, these are also debatable assumptions.

[10]There have been many recent attempts in the educational psychology literature at analyzing selectivity bias in the evaluation model. Among these are Cronbach et al. (1977), Kenny (1975), Linn and Werts (1977), Overall and Woodward (1977), Porter and Chibucos (1975), Rubin (1974), Weisberg (1978). It now appears to us that these attempts all ran astray precisely because they focus on the linear regression (or ANCOVA) of y on $\underline{x}$, z, as indeed we did in our earlier papers: Goldberger (1972), Barnow (1973), Cain (1975). Our use of the t-variable in the present formulation was stimulated by some ideas in Cronbach, et al. (1977) and Weisberg (1978). It is worth noting that the present approach requires no ex ante specification of the direction of selection--"creaming" vs. "scraping". The direction can be inferred ex post from the signs of the coefficients.

APPENDIX

Ashenfelter, O., and Blum, J. 1976. Evaluating the labor market effects of social programs. Princeton, N.J.: Princeton University Industrial Relations Section.

Cano, F. G. 1971. Readings in evaluation research. New York: Russell Sage.

Dye, T. R. 1971. The measurement of policy impact. Florida State University Press.

Evaluation Studies: Review Annual. Vol. 1, 1976, ed. G. B. Glass. Beverly Hills: Sage Publications.

------. Vol. 2, 1977, ed. M. Guttentag with S. Saar. Beverly Hills: Sage Publications.

Guttentag, M. and Streuning, E. L. 1975. Handbook of evaluation research. Vols. I and II. Beverly Hills: Sage Publications.

Haveman, R., and Margolis, J. 1976. Public expenditures and policy analysis. Rev. ed. Chicago: Markham.

Issac, S., and Michael, W. D. 1971. Handbook of research and evaluation. San Diego: Knopp.

James, D. B. 1975. Analyzing poverty policy. Lexington, Mass.: D. C. Heath.

Rossi, P. H. and Williams, W. 1972. Evaluating social action programs: Theory, practice, and politics. New York: Seminar Press.

Scioli, F. P., Jr. and Cook, T. S. 1977. Methodologies for analyzing public policies. Lexington, Mass.: D.C. Heath.

Stanley, J. C. ed. 1967. Improving experimental design and statistical analysis. Chicago: Rand McNally.

Suchman, E. A. 1967. Evaluative research: Principles and practices in public service and social action programs. New York: Russell Sage.

Weiss, C. H. 1972. Evaluating action programs: Readings in social action and education. Boston: Allyn and Bacon.

------. 1972. Evaluation research: Methods of assessing program effectiveness. Englewood Cliffs, N.J.: Prentice-Hall.

Wholey, J., et al. 1970. Federal evaluation policy. Washington, D.C.: Urban Institute.

REFERENCES

Amemiya, T. 1978. The estimation of a simultaneous-equation generalized probit model, Econometrica, 46, 1193-1205.

Ashenfelter, O. 1978. Estimating the effect of training programs on earnings, The Review of Economics and Statistics, 60, 47-57.

------. 1972. Racial discrimination and trade unionism, The Journal of Political Economy, 80, 435-464.

Ashenfelter, O. and Johnson, G. E. 1972. Unionism, relative wages, and labor quality in U.S. manufacturing industries, International Economic Review, 13, 488-508.

Barnow, B. S. 1973. The effects of Head Start and socioeconomic status on cognitive development of disadvantaged children. University of Wisconsin; unpub. Ph.D. dissertation.

Barnow, B. S. and Cain, G. G. 1977. A reanalysis of the effect of Head Start on cognitive development: Methodology and empirical findings, Journal of Human Resources, 12, 177-197.

Ben-Porath, Y. 1973. Labor force participation rates and the supply of labor, The Journal of Political Economy, 81, 697-704.

Cain, G. G. 1975. Regression and selection models to improve nonexperimental comparisons. In C. A. Bennett and A. A. Lumodaine, eds., Evaluation and experiment: Some critical issues in assessing social programs. New York: Academic Press. 297-317.

Campbell, D. T. 1969. Reforms as experiments, American Psychologist, April, pp. 409-429; reprinted with some revisions in E. L. Struening and M. Guttentag, eds. Handbook of Evaluation Research, vol. I. Beverly Hills: Sage Publications, 1975. 71-99.

Campbell, D. T. and Erlebacher, A. 1970. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth, ed., Compensatory education: A national debate, Vol. 3, Disadvantaged child. New York: Brunner/Mazel Publishers. 185-210.

Crawford, D. L. 1979. Estimating models of earnings from truncated samples. University of Wisconsin, unpub. Ph.D. dissertation.

Cronbach, L. J., Rogosa, D. R., Floden, R. E., & Price, G. G. 1977. Analysis of covariance in nonrandomized experiments: Parameters affecting bias, Stanford Evaluation Consortium, Occasional Paper, August.

Deming, W. E. 1975. The logic of evaluation. In E. L. Streuning and M. Guttentag, eds. Handbook of evaluation research, vol. I, Beverly Hills. Sage Publications. 53-68.

Garfinkel, I. 1977. Effects of welfare programs on experimental responses. In H. W. Watts and A. Rees, eds., The New Jersey income-maintenance experiment, vol. III. 279-302.

Garfinkel, I. and Gramlich, E. M. 1973. A statistical analysis of the OEO experiment in educational performance contracting, Journal of Human Resources, 8, 275-305.

Goldberger, A. S. 1972. Selection bias in evaluating treatment effects: Some formal illustrations. Institute for Research on Poverty, University of Wisconsin, Madison, Discussion Paper 123-72.

Greenberg, D. H. and Kosters, M. 1973. Income guarantees and the working poor: The effect of income-maintenance programs on the hours of work of male family heads. In G. G. Cain and H. W. Watts, eds., Income maintenance and labor supply Chicago: Rand McNally Co. 14-101.

Griliches, Z. 1976. Wages of very young men, The Journal of Political Economy, 84, S69-S86.

Gronau, R. 1973. The effect of children on the housewife's value of time, The Journal of Political Economy, 81, S168-S199.

------. 1974. Wage comparisons -- a selectivity bias, The Journal of Political Economy, 82, 1119-1144.

Hanoch, G. 1980. A multivariate model of labor supply: Methodology and estimation. In J. P. Smith, ed., Female labor supply: Theory and estimation. Princeton, N.J. Princeton University Press, 249-326.

Heckman, J. J. 1974. Shadow prices, market wages, and labor supply, Econometrica, 42, 679-694.

------. 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, 475-492.

------. 1978. Dummy endogenous variables in a simultaneous equation system, Econometrica, 46, 931-960.

------. 1979. Sample bias as a specification error, Econometrica, 47, 153-162.

Heckman, J. J. and Wolpin, K. I. 1976. Does the contract compliance program work? An analysis of Chicago data, Industrial and Labor Relations Review, 29, 544-564.

Johnson, N. A. and Kotz, S. 1970. Distributions in statistics: Continuous univariate distributions-1. Boston: Houghton Mifflin.

------. 1972. Distributions in statistics: Continuous multivariate distributions. New York: Wiley.

Kenny, D. A. 1975. A quasi-experimental approach to assessing treatment effects in the nonequivalent control group design, Psychological Bulletin, 82, 345-361.

Lee, L. F. 1978. Unionism and wage rates: A simultaneous equations model with qualitative and limited dependent variables, International Economic Review, 19, 415-433.

------. 1979. Identification and estimation in binary choice models with limited (censored) dependent variables, Econometrica, 47, 977-996.

Leigh, D. E. 1978. An analysis of the interrelation between unions, race, and wage and nonwage compensation. Final report for the U.S. Department of Labor.

Lewis, H. 1959. Competitive and monopoly unionism. In P. D. Bradley, ed., The public stake in union power. Charlottesville, Va.: University of Virginia Press. 181-208.

------. 1974. Comments on selectivity biases in wage comparisons, The Journal of Political Economy, 82, 1145-1156.

Linn, R. L. and Werts, C. E. 1970. Analysis implications of the choice of a structural model in the nonequivalent control group design, Psychological Bulletin, 84, 229-234.

Maddala, G. S. 1976. Self-selectivity problems in econometric models. Department of Economics, University of Florida.

------. 1978. Selectivity problems in longitudinal data, Annales de L'INSEE, 30-31, 423-450.

Maddala, G. S. and Lee, L. F. 1976. Recursive models with qualitative endogenous variables, Annals of Economic and Social Measurement, 5, 525-545.

Olsen, R. J. 1978. Comment on "The effect of unions on earnings and earnings on unions: A mixed logit approach," International Economic Review, 19, 259-261.

Overall, J. E. and Woodward, J. A. 1977. Nonrandom assignment and the analysis of covariance, Psychological Bulletin, 84, 588-594.

Porter, A. C. and Chibucos, T. R. 1975. Common problems of design and analysis in evaluative research, Sociological Methods and Research, 3, 235-257.

Rivlin, A. 1971. Systematic thinking for social action. Washington, D.C.: The Brookings Institution.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies, Journal of Educational Psychology, 66, 688-701.

Schmidt, P. and Strauss, R. P. 1976. The effect of unions on earnings and earnings on unions: A mixed logit approach, International Economic Review, 17, 204-212.

Schultz, T. P. 1980. Estimating labor supply functions for married women. In J. P. Smith, ed., Female labor supply: Theory and estimation. Princeton, N.J.: Princeton University Press. 25-89.

Smith, R. S. 1975. The estimated impact on injuries of OSHA's target industries program. Paper presented at the ASPER-OSHA Conference on Evaluating the Effects of the Occupational Safety and Health Program, March.

Triplett, J. E. 1975. On the methodology of evaluating economic effects of government programs: A comment on Professor Smith's paper. U.S. Department of Labor, BLS Working Paper 41, May.

Waldman, D. M. 1979. Time allocation of young men. University of Wisconsin; unpub. Ph. D. dissertation.

Watts, H. W., Peck, J. K., and Taussig, M. 1977. Site selection, representativeness of the sample, and possible attrition bias. In H. W. Watts and A. Rees, eds., The New Jersey income-maintenance experiment, Vol. III. New York: Academic Press. 441-466.

Weisberg, H. I. 1978. Statistical adjustments and uncontrolled studies. The Huron Institute, unpublished manuscript.