NBER TECHNICAL WORKING PAPER SERIES

RANDOMIZATION AND SOCIAL POLICY EVALUATION

James J. Heckman

Technical Working Paper No. 107

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 1991

This research was supported by NSF Grant SES 87-39151. I have benefited from the comments of Ricardo Barros, Fred Doolittle, Sherry Glied, Joe Hotz, Tom MaCurdy, Charles Manski and James Walker. For a more complete statement of the argument in this paper see Heckman (1990b). This paper is part of NBER's research program in Labor Studies. Any opinions expressed are those of the author and not those of the National Bureau of Economic Research.

NBER Technical Working Paper #107
July 1991

RANDOMIZATION AND SOCIAL POLICY EVALUATION

ABSTRACT

This paper considers the recent case for randomized social experimentation and contrasts it with older cases for social experimentation. The recent case eschews behavioral models, assumes that certain mean differences in outcomes are the parameters of interest to evaluators and assumes that randomization does not disrupt the social program being analyzed. Conditions under which program disruption effects are of no consequence are presented. Even in the absence of randomization bias, ideal experimental data cannot estimate median (other quantile) differences between treated and untreated persons without invoking supplementary statistical assumptions. The recent case for randomized experimentation does not address the choice of the appropriate stage in a multistage program at which randomization should be conducted. Evidence on randomization bias is presented.

James J. Heckman
University of Chicago
Department of Economics
Harris School
1126 E. 59th Street
Chicago, IL 60637
and NBER

This paper considers the benefits and limitations of randomized social experimentation as a tool for evaluating social programs.[1] The argument for social experimentation is by now familiar. Available cross–section and time–series data often possess insufficient variability in critical explanatory variables to enable analysts to develop convincing estimates of the impacts of social programs on target outcome variables. By collecting data to induce more variation in the explanatory variables, more precise estimates of policy impacts are possible. In addition, controlled variation in explanatory variables can make endogenous variables exogenous i.e it can induce independent variation in observed variables relative to unobserved variables. Social experiments induce variation by controlling the way data are collected. Randomization is one way to induce extra variation but it is by no means the only way or even necessarily the best way to achieve the desired variation.

The original case for social experimentation took as its point of departure the Haavelmo (1944) – Marschak (1953) – Tinbergen (1956) social planning paradigm. Social science knowledge was thought to be sufficiently far advanced to be able to identify basic behavioral relationships which, when estimated, could be used to evaluate the impacts of a whole host of social programs, none of which had actually been implemented at the time of the evaluation. The "structural equation" approach to social policy evaluation promised to enable analysts to simulate a wide array of counterfactuals that could be the basis for "optimal" social policy making. The goal of social experimentation as envisioned by Conlisk and Watts (1969) and Conlisk (1973) was to develop better estimates of the

[1]Throughout this paper, I refrain from restating familiar arguments about the limitations of social experiments and focus on a problem not treated in the literature on this topic. See Cook and Campbell (1979), the papers in Hausman and Wise (1983), and the papers in this volume for statements problems of attrition, spillover effects, etc..

structural equations needed to perform the simulation of counterfactuals.

The original proponents of the experimental method in economics focused on the inability of cross–section studies of labor supply to isolate "income" and "substitution" effects needed to estimate the impact of negative income taxes (NIT) on labor supply. Experiments were designed to induce greater variation in wages and incomes across persons to afford better estimation of critical policy parameters. The original goal of these experiments was not to evaluate a specific set of NIT programs but to estimate parameters that could be used to assess the impacts of those and many other possible programs.

As the NIT experiments were implemented, their administrators began to expect less from them. Attention focused on evaluations of specific treatment effects actually in place. (See Cain, 1975). Extrapolation from and interpolation between the estimated treatment effects took the place of counterfactual policy simulations based on estimated structural parameters as the method of choice for evaluating proposed programs not actually implemented. (See Hausman and Wise, 1983).

The recent case for randomized social experiments represents a dramatic retreat from the ambitious program of "optimal" social policy analysis that was never fully embraced by most economists and was never embraced by other social scientists. Considerable skepticism has recently been expressed about the value of econometric or statistical methods for estimating the impacts of specific social programs or the parameters of "structural" equations required to simulate social programs not yet in place. Influential studies by LaLonde (1986) and Fraker and Maynard (1987) have convinced many that econometric and statistical methods are incapable of estimating true program impacts from non–randomized data.

Recent advocates of social experiments are more modest in their ambitions than were the original proponents. They propose to use randomization to evaluate programs actually in place (be they ongoing programs or pilot "demonstration" projects) and to avoid invoking the litany of often unconvincing assumptions that underlie "structural" or

"econometric" or "statistical" approaches to program evaluation.[2] Their case for randomization is powerfully simple and convincing: randomly assign persons to a program and compare target responses of participants to those of randomized–out nonparticipants. The mean difference between participants and randomized–out nonparticipants is defined to be the effect of the program. Pursuit of "deep structural" parameters is abandoned. No elaborate statistical adjustments or arbitrary assumptions about functional forms of estimating equations are required to estimate the parameter of interest using randomized data. No complicated estimation strategy is required. Everyone understands means. Randomization ensures that there is no selection bias among participants i.e., there is no selection into or out of the program on the basis of outcomes for the randomized sample.

Proponents of randomized social experiments implicitly make an important assumption: that randomization does not alter the program being studied. For certain evaluation problems and for certain behavioral models this assumption is either valid or innocuous. For other problems and models it is not. A major conclusion of this study is that advocates of randomization have overstated their case for having avoided arbitrary assumptions. Evaluation by randomization makes implicit behavioral assumptions that in certain contexts are quite strong. Bias induced by randomization is a serious possibility.

In addition, advocates of randomization implicitly assume that certain mean differences in outcomes are invariably the object of interest in performing an evaluation. In fact, experimental methods cannot estimate median differences without invoking stronger assumptions than are required to recover means. There are many parameters of potential interest, only some of which can be cast into a mean difference framework. The parameters of interest may not be defined by a hypothetical randomization and randomized data may not be ideal for estimating these parameters.

Advocates of randomization are often silent on an important practical matter.

---

[2]In an early contribution Orcutt and Orcutt (1968) suggest this use of social experiments.

Many social programs are multistage in nature. At what stage should randomization occur? At the enrollment, assignment to treatment, promotion, review of performance, or placement stage? The answer to this question reveals a contradiction in the case for randomized experiments. In order to use simple methods (i.e., mean differences between participants and nonparticipants) to evaluate the effects of the various stages of a multistage program, it is necessary to randomize at each stage. Such multistage randomization has never been implemented probably because it would drastically alter the program being evaluated. But if only one randomization can be conducted, an evaluation of all stages of a multistage program entails the use of the very controversial nonexperimental methodology sought to be avoided in the recent case for social experimentation.

The purpose of this paper is to clarify the arguments for and against randomized social experiments. In order to focus the discussion, I first present (in section 1) a prototypical social program and consider what features of the program are of interest to policy evaluators. In Section 2, I discuss the difficulties that arise in determining program features of interest. A precise statement of the evaluation problem is given. In Section 3, I state the case for simple randomization. Then I consider the implicit behavioral assumptions that underlie the case and conditions under which they hold. I also discuss what parameters can and cannot be recovered from randomized social experiments even under ideal conditions. The case for social experiments assumes that certain means are of paramount interest. Experiments are much less effective in recovering medians or more general features of distributions. In Section 4, I present some indirect evidence on the validity of these assumptions for the case of a recent evaluation of the Job Training Parternship Act (JTPA). I also consider some parallel studies of their validity in the randomized clinical trials literature in medicine. In Section 5, I discuss the issue of choosing the appropriate stage at which one should randomize in a multistage program. In Section 6, I discuss the tension between the recent and the older cases for social

experimentation. The final section summarizes the argument.

1. The Questions of Interest in Evaluating A Prototypical Social Program

In order to focus ideas, it is helpful to consider the evaluation of a prototypical social program. The prototype considered here is a manpower training program similar to the JTPA program described by Hotz in his paper in this volume.

The prototypical program offers a menu of training options to potential trainees. Specific job–related skills may be learned as well as general skills (i.e., reading, writing and arithmetic). Remedial general training may precede specific training. Job placement may be offered as a separate service independently of any skill acquisition or after completion of such activity. Some specific skill programs entail working for an employer at a subsidized wage (i.e. , on the job training).

Individuals who receive training proceed through the following steps: (1) they apply; (2) are accepted; (3) are placed in a specific training sequence ; (4) are reviewed; (5) are certified in a skill; and (6) are placed with an employer. For trainees receiving on–the–job training, steps (3)–(6) are combined although trainees may be periodically reviewed during their training period. Individuals may drop out or be rejected at each stage.

Training centers are paid by the U.S. Government based on the quality of the placement of their trainees. Quality is measured by the wages received over a specified period of time after trainees complete their training program (e.g., six months). Managers thus have an incentive to train persons who are likely to be high quality placements and who can achieve that status at low cost to the center. Trainees receive compensation (subsidies) while in the program. Training centers recruit trainees through a variety of promotional schemes.

There are many questions of interest to program evaluators. The question that receives the most attention is the effect of training on the trained:

Q–1: "*What is the effect of training on the trained*?"

This is the "bottom line" stressed in many evaluations. When costs of a program are subtracted from the answer to *Q–1*, and returns are appropriately discounted, the net benefit of the program is produced for a fixed group of trainees.

But there are many other questions that are also of potential interest to program evaluators such as:

Q–2: "*What is the effect of training on randomly assigned trainees?*"

The answer to *Q–2* would be of great interest if training were mandated for an entire population as in workfare programs that force welfare recipients to take training. Another question of interest concerns application decisions:

Q–3: "*What is the effect of subsidies (and/or advertising, and/or local labor market conditions, and/or family income, and/or race, sex) on application decisions?*"

There are many other questions of potential interest such as:

Q–4: "*What are the effects of center performance standards, profit rates, local labor market structure, and governmental monitoring on training center acceptance of applicant decisions and placement in specific programs?*"

and

Q–5: "*What are the effects of family background, center profit rates, subsidies and local labor market conditions on the decision to dropout from a program and the length of time taken to complete the program?*"

and

Q–6: "*What are the effects of labor market conditions, subsidies, profit rates, <u>etc.</u>, on placement rates and wage and hour levels attained at placement?*"

and

Q–7: "*What is the cost of training a worker in the various possible ways?*"

Answers to all of these questions and refinements of them are of potential interest to policy makers. The central evaluation problem is how to obtain convincing answers to them.

2. The Evaluation Problem

To characterize the essential features of the evaluation problem, it is helpful to concentrate on only a few of the questions listed above. I focus attention on questions 1 and 2 and a combination of the ingredients in questions *Q–3* and *Q–4*:

Q–3' "*What are the effects of the variables listed in Q–3 and Q–4 on application and enrollment of individuals?*"

To simplify the analysis I assume throughout the discussion in this section that there is only one type of treatment administered by the program, so determining assignment to treatment is not an issue. I assume no attrition from the program and that length of participation in the program is fixed. These assumptions would be true if, for example, the ideal program occurs at a single instant in time and gives every participant the same "dose" although the response to the dose may differ across people. I also assume absence of any interdependence among units resulting from common, site–specific unobservables or feedback effects.[3]

This paper does not focus exclusively or even mainly on "structural estimation" because it is not advocated in the recent literature on social experiments and because a discussion of that topic raises additional issues not germane to this paper. Structural

[3]This is Rubin's "SUTVA" assumption. See Holland (1986). It is widely invoked in the literature in econometrics and statistics even though it is often patently false. See Garfinkel, Manski and Michaelopolous (this volume) for a discussion of this problem.

approaches require specification of a common set of characteristics and a model of program participation and outcomes to describe all programs of potential interest. They require estimating responses to variations in characteristics that describe programs not yet put in place. This requires specification and measurement of a common set of characteristics that underlie such programs.

The prototypical structural approach is well illustrated in the early work on estimating labor supply responses to negative income tax programs. Those programs operated by changing the wage level and income level of potential participants. Invoking the neoclassical theory of labor supply, if one can determine the response of labor supply to changes in wages and income levels, (the "substitution" and "income" effects respectively), one can also determine who would participate in a program (see, e.g., Ashenfelter, 1983). Thus from a common set of parameters one can simulate the effect of all possible NIT programs on labor supply.

It is for this reason that early advocates of social experiments sought to design experiments that would give maximal sample independent variation in wage and income levels across subjects so that precise estimates of wage and income effects could be obtained. Cain and Watts (1974) argued that in cross–section data, wages and income were sufficiently highly correlated and the variability in sample incomes was sufficiently small that it was difficult, if not impossible, to estimate separate wage and income effects on labor supply.

The structural approach is very appealing when it is credible. It focuses on essential aspects of responses to programs. But its use in practice requires invoking strong behavioral assumptions in order to place diverse programs on a common basis. In addition, it requires that the common characteristics of programs can be measured. Both the measurement problems and the behavioral assumptions required in the structural approach raise issues outside the scope of this paper. I confine most of my attention to the practical—and still very difficult—problem of evaluating the effect of existing programs

and the responses to changes in parameters of those programs that might affect program participation.

(a) A Model of Program Evaluation

To be more specific, it is useful to define variable $D = 1$ if a person participates in a hypothetical program. $D = 0$ otherwise. If a person participates, she/he receives outcome Y_1. Otherwise she/he receives Y_0. Thus the observed outcome Y is:

(1) $$Y = Y_1 \text{ if } D = 1$$
$$Y = Y_0 \text{ if } D = 0.$$

A crucial feature of the evaluation problem is that we do not observe the same person in both states. This is called "the problem of causal inference" by some statisticians (see, e.g., Holland, 1986). Let Y_1 and Y_0 be determined by X_1 and X_0 respectively. Presumably X_1 includes relevant aspects of the training received by trainees. X_0 and X_1 may contain background and local labor market variables. We write functions relating those variables to Y_0 and Y_1 respectively:

(2a) $$Y_1 = g_1(X_1)$$

(2b) $$Y_0 = g_0(X_0).$$

In terms of more familiar linear equations, (2a) and (2b) may be specialized to

(2a)' $$Y_1 = X_1\beta_1$$

and

(2b)' $$Y_0 = X_0\beta_0$$

respectively.

Let Z be variables determining program participation. If

(3) $$Z \in \Psi,\ D = 1;\ Z \notin \Psi,\ D = 0$$

where Ψ is a subset of the possible Z values. If persons have characteristics that lie in set Ψ, they participate in the program. Otherwise they do not. Included among the Z are characteristics of persons and their labor market opportunities as well as characteristics of

the training sites selecting applicants. In order to economize on symbols, I represent the entire collection of explanatory variables by $C = (X_0, X_1, Z)$. If some variable in C does not appear in X_1 or X_0, its coefficient or associated derivative in g_1 or g_0 is set to zero for all values of the variable.

If one could observe all of the components of C for each person in a sample, one might still not be able to determine g_1, g_0 and Ψ. The available samples might not contain sufficient variation in the components of these vectors to trace out g_0, g_1 or to identify set Ψ. Recall that it was a "multicollinearity" problem (in income and wage variables needed to determine labor supply equations) and a lack of sample variation in income that partly motivated the original proponents of social experiments in economics.

Assuming sufficient variability in the components of the explanatory variables, one can utilize data on participants to determine g_1, on nonparticipants to determine g_0, and the combined sample to determine Ψ. With knowledge of these functions and sets, one can readily answer evaluation problems *Q–1*, *Q–2* and *Q–3*.[4] It would thus be possible to construct Y_1 and Y_0 for each person and to estimate the gross gain to participation for each participant or for each person in the sample. In this way questions *Q–1* and *Q–2* can be fully answered. From knowledge of Ψ it is possible to fully answer question *Q–3* for each person.

As a practical matter, analysts do not observe all of the components of C. The unobserved components of these outcome and enrollment functions are a major source of evaluation problems. It is these missing components that motivate treating Y_1, Y_0 and D as random variables, conditional on the available information. This intrinsic randomness rules out a strategy of determining Y_1 and Y_0 for each person. Instead, a statistical approach is adopted that focuses on estimating the joint distribution of Y_1, Y_0, D conditional on the available information or some features of it.

---

[4]Provided that the support of the X_1, X_0 and Z variables in the sample covers the support of these variables in the target populations of interest.

Let subscript "i" denote available information. Thus C_i contains the variables available to the analyst thought to be legitimate for determining Y_1, Y_0 and D. These variables may consist of some components of C as well as proxies for the missing components.

The joint distribution of Y_1, Y_0, D given $C_i = c_i$ is

$$F(y_0, y_1, d|c_i) = \Pr(Y_0 \leq y_0, Y_1 \leq y_1, D = d|C_i = c_i) \tag{4}$$

where I follow convention by denoting random variables by upper case letters and their realization by lower case letters. If (4) can be determined, and the distribution of C_i is known, it is possible to answers questions *Q–1*, *Q–2* and *Q–3'* in the following sense: one can determine population distributions of Y_0, Y_1 and the population distribution of the gross gain from program participation

$$\Delta = Y_1 - Y_0$$

and one can write out the probability of the event $D = 1$ given Z_i.

(b) The Parameters of Interest in Program Evaluation

We can answer *Q–1* from knowledge of

$$F(y_0, y_1|D = 1, c_i)$$

and hence

$$F(\delta|D = 1, c_i)$$

(the distribution of the effect of treatment on the treated where δ is the lower case version of Δ). One can answer *Q–2* by determining

$$F(y_0, y_1|c_i)$$

which can be produced from (4) and the distribution of the explanatory variables by elementary probability operations. In this sense one can determine the gains from randomly moving a person from one distribution $(F(y_0|c_i))$ to another $(F(y_1|c_i))$. The answer to *Q–3'* can be achieved by computing the probability of participation:

$$\Pr(D = 1|c_i) = F(d|c_i)$$

from (4).

In practice, comparisons of means occupy most of the attention in the literature, although medians, or other quantiles, are also of interest. The means are assumed to exist. Much of the literature defines the answer to *Q–1* as

$$E(\Delta \mid D = 1, c_i) = E(Y_1 - Y_0 \mid D = 1, c_i) \tag{6}$$

and the answer to *Q–2* as

$$E(\Delta \mid c_i) = E(Y_1 - Y_0 \mid c_i) \tag{7}$$

although in principle knowledge of the full distribution of Δ, or some other feature besides the mean (e.g., the median), might be desirable.

Even if the means in (6) or (7) were zero, it is of interest to know what fraction of participants or the population would benefit from a program. This would require knowledge of $F(\delta \mid D = 1, c_i)$ or $F(\delta \mid c_i)$ respectively. In order to ascertain the existence of "cream skimming" (i.e, defined in one version of that concept as the phenomenon that training sites select the best people into a program–those with high values of Y_0 and Y_1), it is necessary to know the correlation or stochastic dependence between Y_1 and Y_0. This would require knowledge of features of

$$F(y_1, y_0 \mid D = 1, c_i)$$

or

$$F(y_1, y_0 \mid c_i)$$

other than the means of Y_1 and Y_0. To answer many questions, knowledge of mean differences is inadequate or incomplete.

Determining the joint distribution (4) is a difficult problem. In the next section I show that randomized social experiments of the sort proposed in the recent literature do not produce data sufficient for this task.

The data routinely produced from social program records enable analysts to determine

$$F(y_1 \mid D = 1, c_i),$$

the distribution of outcomes for participants, and

$$F(y_0 | D = 0, c_i)$$

the distribution of outcomes for nonparticipants, and they are sometimes sufficiently rich to determine

$$\Pr(D = 1 | c_i) = F(d | c_i),$$

the probability of participation. But unless further information is available, these pieces of information do not suffice to determine (4). By virtue of (1), there are no data on both components of (Y_1, Y_0) for the same person. In general, for the same values of $C_i = c_i$,

8(a) $$F(y_0 | D = 1, c_i) \neq F(y_0 | D = 0, c_i)$$

and

8(b) $$F(y_1 | D = 1, c_i) \neq F(y_1 | D = 0, c_i)$$

which gives rise to the problem of selection bias in the outcome distributions. The more common statement of the selection problem is in terms of means:

9(a) $$E(\Delta | D = 1, c_i) \neq E(Y_1 | D = 1, c_i) - E(Y_0 | D = 0, c_i)$$

9(b) $$E(\Delta | c_i) \neq E(Y_1 | c_i) - E(Y_0 | c_i)$$

i.e., persons who participate in a program are different people from persons who do not participate in the sense that the mean outcomes of participants in the nonparticipation state would be different from those of non–participants even after adjusting for C_i.

A whole host of alternative methods has been proposed for solving the selection problem either for means or for entire distributions. Heckman and Honore' (1990), Heckman and Robb (1985, 1986), Heckman (1990a,b), and Manski (1990) offer alternative comprehensive treatments of the various approaches to this problem in econometrics and statistics. Some untestable apriori assumption must be invoked to recover the missing components of the distribution. Constructing these counterfactuals inevitably generates

controversy.[6]

LaLonde (1986) and Fraker and Maynard (1987) have argued that these controversies are of more than academic interest. In influential work analyzing randomized experimental data using non–experimental methods, these authors produce a wide array of estimates of impacts of the same program using different nonexperimental methods. They claim that there is no way to choose among competing nonexperimental estimators.

Heckman and Hotz (1989) reanalyze their data and demonstrate that their claims are greatly exaggerated. Neither set of authors performed model specification tests. When such tests are performed, they eliminate all but the nonexperimental models that reproduce the inference obtained by experimental methods.

There is, nonetheless, a kernal of truth in the criticism of LaLonde (1986) and Fraker and Maynard (1987). Each test proposed by Heckman and Hotz (1989) has its limitations. Tests of overidentifying features of a model can be rendered worthless by changing the model to a just–identified form. Tests that check if nonexperimental selection bias methods adjust for preprogram differences in outcome measures have no power against the alternative hypothesis that selection occurs on post–program differences between participants and nonparticipants that are stochastically independent of pre–program differences. (Heckman 1990b).

All nonexperimental methods are based on some maintained, untestable, assumption. The great source of appeal of randomized experiments is that they appear to require no assumptions. In the next section I demonstrate that the case for randomized evaluations rests on unstated assumptions about the problem of interest, the number of stages in a program and the response of agents to randomization. These assumptions are different from and not arguably better than the assumptions maintained in the

[6]Manski (1990) has shown that it is sometimes possible to bound $E(\Delta \mid D = 1, c_i)$ and $E(\Delta \mid c_i)$ even if they cannot be determined exactly.

nonexperimental econometrics and statistics literature.

3. The Case For And Against Randomized Social Experiments

The case for randomized social experiments is almost always stated within the context of obtaining answers to questions *Q–1* and *Q–2* — the "causal problem" as defined by statisticians. (See Fisher (1935), Cox (1958), Rubin (1978) and Holland (1986)). From this vantage point, the participation equation that answers *Q–3'* is a "nuisance function" that may give rise to a selection problem. Simple randomization makes treatment status statistically independent of (Y_1, Y_0, C).

To state the case for randomization most clearly, it is useful to introduce a variable A indicating actual participation in a program:

$$A = 1 \text{ if a person participates}$$
$$= 0 \text{ otherwise}$$

and separate it from variable D indicating who would have participated in a program in a non–experimental regime. Let D^* denote a variable indicating if an agent is at risk for randomization (i.e., if the agent applied and was accepted in a regime of random selection):

$$D^* = 1 \text{ if a person is at risk for randomization}$$
$$= 0 \text{ otherwise.}$$

In the standard approach, randomization is implemented at a stage when D^* is revealed. Given $D^* = 1$, A is assumed independent of (Y_0, Y_1, C) so

$$F(y_0, y_1, c, a | D^* = 1) = F(y_0, y_1, c | D^* = 1)F(a | D^* = 1).$$

More elaborate randomization schemes might be implemented but are rarely proposed.

Changing the program enrollment process by randomly denying access to individuals who apply and are deemed suitable for a program may make the distribution of D^* different from D. Such randomization alters the information set of potential applicants and program administrators unless neither is informed about the possibility of

randomization – an unlikely event for an ongoing program or for one–shot programs in many countries such as the U.S. where full disclosure of program operating rules is required by law. Even if it were possible to surprise potential trainees, it would not be possible to to surprise training centers administering the program. (Recall that D^* is the outcome of joint decisions by potential trainees and training centers.) The conditioning set determining D^* differs from that of D by the inclusion of the probability of selection ($p = Pr(A = 1)$), i.e. it includes the effect of randomization on agent and center choices.

Proponents of randomization invoke the assumption that:

AS–1: $$Pr(D = 1|c) = Pr(D^* = 1|c,p)$$

or assume that it is "practically" true.[6]

There are many reasons to suspect the validity of this assumption. If individuals who might have enrolled in a nonrandomized regime make plans anticipating enrollment in training, adding uncertainty at the acceptance stage may alter their decision to apply or to undertake activities complementary to training. Risk averse persons will tend to be eliminated from the program. Even if randomization raises agent utility[7], behavior will be altered. If training centers must randomize after a screening process, it might be necessary for them to screen more persons in order to reach their performance goals and this may result in lowered trainee quality. Degradation in the quality of applicants might arise even if slots in a program are rationed. Randomization may solve rationing problems in an equitable way if there is a queue for entrance into the program but it may also alter the composition of the trainee pool.

Assumption AS–1 is entirely natural in the context of agricultural and biological experimentation in which the Fisher model of randomized experiments was originally

---

[6]Failure of this assumption is an instance of the Lucas critique (1976) applied to social experimentation. It is also an instance of a "Hawthorne" effect. See Cook and Campbell (1979).

[7]This can arise even if agents are risk–averse by convexifying a non–convex problem. See Arnott and Stiglitz (1988).

developed. However, the Fisher model is a potentially misleading paradigm for social science. Humans act purposively and their behavior is likely to be altered by introducing randomization into their choice environment. The Fisher model may be ideal for the study of fertilizer treatments on crop yields. Plots of ground do not respond to anticipated treatments of fertilizer nor can they excuse themselves from being treated. Commercial manufacturers of fertilizer can be excluded from selecting favorable plots of ground in an agricultural experimental setting in a way that training center managers cannot be excluded from selecting favorable trainees in a social science setting.

If AS–1 is true,

10(a) $$F(y_1,c|A=1) = F(y_1,c|D^* = 1) = F(y_1,c|D=1)$$

10(b) $$F(y_0,c|A=0) = F(y_0,c|D^* = 1) = F(y_0,c|D=1).$$

(11) $$E(Y_1|A=1) - E(Y_0|A=0) = E(\Delta|D=1).$$

Simple mean difference estimators between participants and randomized–out–non–participants answer question *Q–1* stated in terms of means, at least for large samples. The distribution of explanatory variables C is the same in samples conditioned on A. The samples conditioned on A = 1 and A = 0 are thus balanced.

In this sense, randomized data are "ideal". People untrained in statistics, such as politicians and program administrators, understand means, and no elaborate statistical adjustments or functional form assumptions about a model are imposed on the data. Moreover (11) may be true even if AS–1 is false.

This is so for the widely used dummy endogenous variable model. (Heckman, 1978). For that case

(12) $$Y_1 = \alpha + Y_0.$$

This model is termed the "fixed treatment effect for all units model" in the statistics literature. (See Cox (1958)). That model writes

$$Y_1 = g_1(x_1) = \alpha + g_0(x_0) = \alpha + Y_0$$

so the effect of treatment is the same for everyone. In terms of the linear regression model

of (2a)′ and (2b)′ this model can be written as $X_1\beta_1 = \alpha + X_0\beta_0$. Even if AS–1 is false, (11) is true because

$$\begin{aligned} & E(Y_1|A=1) - E(Y_0|A=0) \\ & = E(\alpha + Y_0|A=1) - E(Y_0|A=0) \\ & = \alpha + E(Y_0|D^*=1) - E(Y_0|D^*=1) \\ & = \alpha \\ & = E(\Delta|D=1) \\ & = E(\Delta). \end{aligned}$$

The dummy endogenous variable model is widely used in applied work. Reliance on this model strengthens the popular case for randomization. Questions 1 and 2 have the same answer in this model and randomization provides a convincing way to answer both.

The requirement of treatment outcome homogeneity can be weakened and (11) can still be justified if (AS–1) is false. Suppose there is a random response model (sometimes called a random effects model)

13(a) $$Y_1 = Y_0 + (\alpha + \Xi)$$

where Ξ is an individual's idiosyncratic response to treatment after taking out a common response α and

13(b) $$E(\Xi|D^*=1) = 0,$$

then (11) remains true. If potential trainees and training centers do not know the trainees' gain from the program in advance of his/her enrollment in the program, and they use α in place of $\alpha + \Xi$ in making participation decisions, then (11) is still satisfied. Thus even if responses to treatments are heterogeneous, the simple mean – difference estimator obtained from experimental data may still answer the mean difference version of question 1.

It is important to note how limited are the data obtained from an "ideal" social experiment (i.e., one that satisfies AS–1). Without invoking additional assumptions, one cannot estimate the distribution of Δ conditional or unconditional on $D = 1$. One cannot estimate the median of Δ nor can one determine the empirical importance of

"cream–skimming" (the stochastic dependence between Y_0 and Y_1) from the data. One cannot estimate $E(\Delta)$. Both experimental and nonexperimental data are still plagued by the fundamental problem that one cannot observe Y_0 and Y_1 for the same person. Randomized experimental data of the type proposed in the literature only facilitate simple estimation of one parameter

$$E(\Delta \mid D = 1,c).$$

Assumptions must be imposed to produce additional parameters of interest even from ideal experimental data. Answers to most of the questions stated in Section 1 still require statistical procedures with their attendant, controversial assumptions.

If assumption AS–1 is not satisfied, the final equalities in 10(a) and 10(b) are not satisfied and in general

$$E(Y_1 \mid A = 1) - E(Y_0 \mid A = 0)$$
$$\neq E(\Delta \mid D = 1).$$

Moreover, the data produced by the experiment will not enable analysts to assess the determinants of participation in a nonrandomized regime because the application and enrollment decision processes will have been altered by randomization: i.e.

$$Pr(D = 1 \mid c) \neq Pr(D^* = 1 \mid c,p)$$

unless $p = 1$. Thus experimentation will not produce data to answer question *Q–3'* unless randomization is a permanent feature of the program being evaluated.

In the general case in which agent response to programs is heterogenous ($\Xi \neq 0$) and agents anticipate this heterogeneity (more precisely, Ξ is not stochastically independent of D), assumption (AS–1) plays a crucial role in justifying randomized social experiments. While (AS–1) is entirely non–controversial in some areas of science–such as in agricultural experimentation where the original Fisher model was developed–it is more problematic in social settings. It may produce clear answers to the wrong question and may produce data that cannot be used to answer crucial evaluation questions, even when question *Q–1* can be clearly answered.

4. Evidence on Randomization Bias

Violations of assumption (AS–1) in general make the evidence from randomized social experiments unreliable. How important is this theoretical possibility in practice? Surprisingly, very little is known about the answer to this question for the social experiments conducted in economics. This is so because, except for one program, randomized social experimentation has only been implemented on "pilot projects" or "demonstration projects" designed to evaluate new programs without precedent. The possibility of disruption by randomization cannot be confirmed or denied on data from these experiments. In the one ongoing program evaluated by randomization, participation was compulsory for the target population. (Doolittle and Traeger, 1990). Hence randomization did not affect applicant pools or assessments of applicant eligibility by program administrators.

Fortunately there is some information on this question but it is indirect. In response to the wide variability in estimates of the impact of manpower programs derived from non–experimental estimators by LaLonde (1986) and Fraker and Maynard (1987), the U.S. Department of Labor financed a large scale experimental evaluation of the ongoing large scale Job Training Partnership Act (JTPA) which is the main vehicle for providing government training in the U.S. Randomization evaluation was implemented in a variety of sites. The organization implementing this experiment – the Manpower Demonstration Research Corporation (MDRC) – has been an ardent and effective advocate for the use of randomization as a means of evaluating social programs.

A recent report by this organization (Doolittle and Traeger, 1990) gives some indirect information from which it is possible to do a crude revealed preference analysis.[8] Job training in the U.S. is organized through geographically decentralized

[8]Hotz (in this volume) also summarizes their discussion.

centers. These centers receive incentive payments for placing unemployed persons and persons on welfare in "high paying" jobs. The participation of centers in the experiment was not compulsory. Funds were set aside to compensate job centers for the administrative costs of participating in the experiment. The funds set aside range from 5% to 10% of the total operating cost of the centers.

In attempting to enroll geographically dispersed sites MDRC experienced a training center refusal rate in excess of 90%. The reasons for refusal to participate are given in Table 1. (The reasons stated there are not mutually exclusive). Leading the list are ethical and public relations objections to randomization. Major fears (items 2 and 3) were expressed about the effects of randomization on the quality of the applicant pool that would impede the profitability of the training centers. By randomizing, the centers had to widen the available pool of persons deemed eligible and there was great concern about the effects of this widening on applicant quality – precisely the behaviour ruled out by assumption AS–1. In attempting to entice centers to participate, MDRC had to reduce the randomized rejection probability from 1/2 to as low as 1/6 for certain centers. The resulting reduction in the size of the control sample impairs the power of statistical tests designed to test the null hypothesis of no program effect. Compensation was expanded seven–fold to get any centers to participate in the experiment. The MDRC analysts conclude:

> "*Implementing a complex random assignment research design in an ongoing program providing a variety of services does inevitably change its operation in some ways... The most likely difference arising from a random assignment field study of program impacts...is a change in the mix of clients served. Expanded recruitment efforts needed to generate the control group, draw in additional applicants who are not identical to the people previously served. A second likely change is that the treatment categories may somewhat restrict program staff's flexibility to change service recommendations*" (*Doolittle and Traeger, 1990, p. 121*).

Table 1	
Percent of Training Centers Cited Specific Concerns About Participating in The Experiment	
Concern	Percent of Training Centers Citing The Concern
(1) Ethical and Public Relations Implications of:	
(a) Random Assignment in Social Programs	61.8
(b) Denial of Services to Controls	54.4
(2) Potential Negative Effect of Creation of a Control Group on Achievement of Client Recruitment Goals	47.8
(3) Potential Negative Impact on Performance Standards	25.4
(4) Implementation of the Study When Service Providers Do Intake	21.1
(5) Objections of Service Providers to the Study	17.5
(6) Potential Staff Administrative Burden	16.2
(7) Possible Lack of Support by Elected Officials	15.8
(8) Legality of Random Assignment and Possible Grievances	14.5
(9) Procedures for Providing Controls With Referrals to Other Services	14.0
(10) Special Recruitment Problems for Out-of-School Youth	10.5
Sample Size	228

Source: Based on responses of 228 Training Centers contacted about possible participation in the National JTPA Study. (Doolittle and Traeger, 1990, Table 2.1, p. 34).

Notes: Concerns noted by fewer than 5 percent of the training centers are not listed. Percents may add to more than 100.0 because training centers could raise more than one concern.

These authors go on to note that

> *"Some [training centers] because of severe recruitment problems or up–front services cannot implement the type of random assignment model needed to answer the various impact questions without major changes in procedures"* (*Doolittle and Traeger, 1990, p. 123*).

This evidence is indirect. Training centers may offer these arguments only as a means of avoiding administrative scrutiny and there may be no "real" effect of randomization. In a short while data will be available to determine, if in the training centers that did participate in the randomized experiments, center performance declined during the period when randomization was used, or if the mix of trainees in the centers was altered. Self–selection likely guarantees that participant sites are the least likely sites to suffer disruption. Such selective participation in the experiment calls into question the validity of the experimental estimates as a statement about JTPA as a whole. However, we will have a lower bound estimate of the impact of disruption.

Randomization is also controversial in clinical trials analysis in medicine which is sometimes held up as a paragon for empirical social science. (Ashenfelter and Card, 1985). The ethical problem raised by the manpower training centers of denying equally qualified persons access to training has its counterpart in the application of randomized clinical trials. For example, Joseph Palca writing in Science Magazine (1989), notes that AIDS patients denied potentially life–saving drugs take steps to undo random assignment. Patients had the pills they were taking tested to see if they were getting a placebo, or an unsatisfactory treatment, and were likely to drop out of the experiment in either case or seek more effective medication or both. In the MDRC experiment, in some sites qualified trainees found alternative avenues for securing exactly the same training presented by the same subcontractors by using other methods of financial support.

Writing in the Journal of The American Medical Association, Kramer and

Shapiro (1984, p. 2739) note that subjects in drug trials were less likely to participate in randomized trials than in non–experimental studies. They discuss one study of drugs administered to children afflicted with a disease. The study had two components. The non–experimental phase of the study had a four percent refusal rate. Thirty–four percent of a subsample of the same parents refused to participate in a randomized subtrial.

These authors cite evidence suggesting that non–response to randomization is selective. In a study of treatment of adults for cirrhosis, no effect of the treatment was found for participants in a randomized trial. But the death rates for those randomized out of the treatment were substantially lower than among those individuals who refused to participate in the experiment, despite the fact that both groups were administered the same alternative treatment.

This evidence qualifies the case for randomized social experimentation. Where feasible, it may alter the program being studied. For many social programs it is not a feasible tool for evaluation.

5. At What Stage Should Randomization Be Implemented?

Thus far I have deliberately abstracted from the multistage feature of most social programs. In this section I briefly consider the issue of the choice of the stage in a multistage program at which randomization should be implemented.

In principle, randomization could be performed to evaluate outcomes at each stage. The fact that such multiple randomization has never been performed likely indicates that it would exacerbate the problem of randomization bias discussed in Sections 3 and 4. Assuming the absence of randomization bias, if only one randomization is to be performed, at what stage should it be placed?

One obvious answer is at the stage where it is least disruptive although that

stage is not so easy to determine in the absence of considerable information about the process being studied. If randomization is performed at one stage, non–experimental "econometric" or "statistical" estimators are required to evaluate outcomes attributable to participation at all other stages. This accounts for the sometimes very complicated (Ham and LaLonde, 1989) or controversial (Cain and Wissoker, 1990 and Hannan and Tuma, 1990) analyses of randomized experimental data that have appeared in the recent literature.

Moreover, for some of the questions posed in Section 1, it is not obvious that randomization is the method of choice for securing convincing answers. Many of the questions stated in Section 1 concern the response of trainees and training centers to variations in constraints. While enhanced variation in explanatory variables (in a sense made precise by Conlisk (1973)) facilitates estimation of response functions, there is no reason why randomized allocations are desirable or optimal for this purpose.

Thus if we seek to enhance our knowledge of how family income determines program participation, it is not obvious that randomly allocated allotments of family income supplements are a cost effective or optimal substitute for nonexperimental optimal sample design strategies that oversample family incomes at the extremes of the eligible population.[9] If we seek to enhance our knowledge about how local labor market conditions affect enrollment, retention and training–center acceptance and placement decisions, variation across training sites in these conditions would be desireable. It is not obvious that randomization is the best way to secure this variation.

Randomization in eligibility for the program has been proposed as an alternative to randomization at enrollment. This is sometimes deemed to be a more acceptable randomization point because it avoids the application and screening costs

[9]This remark assumes a linear model. For optimal designs in nonlinear models see e.g., Silvey (1980).

that are incurred when accepted individuals are randomized out of a program. Since the randomization is performed outside of the training center, it prevents the training center from bearing the political costs of denying eligible persons the right to participate in the program. For this reason, it is thought to be less disruptive than randomization performed at some other stage.

If eligibility is randomly assigned in the population with probability q and such assignment does not affect the decision to participate in the program among the eligibles, a simple mean difference comparison between treated and untreated persons is <u>less</u> biased for $E(\Delta|D = 1)$ than would be produced from a mean difference comparison between treated and untreated samples without randomized eligibility. In general, the simple mean difference estimator will still be biased. Thus if $p = Pr(D = 1)$ and e denotes eligibility, and $Pr(e = 1) = q$,

$$E(Y_1|D = 1, e = 1) - E(Y_0|e = 0)$$

$$= E(Y_1|D = 1) - \left\{E(Y_0|D=1)\frac{p(1-q)}{p(1-q)+1-p} + E(Y_0|D=0)\frac{(1-p)}{p(1-q)+1-p}\right\}$$

$$= E(Y_1-Y_0|D = 1) + \frac{(1-p)}{p(1-q)+1-p}\left[E(Y_0|D = 1) - E(Y_0|D = 0)\right]$$

$$= E(\Delta|D = 1) + \frac{(1-p)}{p(1-q)+1-p}\left[E(Y_0|D = 1) - E(Y_0|D = 0)\right]$$

so the bias is smaller in absolute value than would be obtained in nonrandomized data:

$$E(Y_1|D = 1) - E(Y_0|D = 0)$$
$$= E(\Delta|D = 1) + \{E(Y_0|D = 1) - E(Y_0|D = 0)\}$$

so long as $0 < q < 1$ (assuming $0 < p < 1$). The intuition is clear: by making some potential participants ineligible, the nonparticipant population now includes some persons whose mean outcomes are the same as what participant outcomes would have

been if they did not participate.

6. The Tension Between The Case For Social Experiments As A Substitute for Behavioral Models and Social Experiments As A Supplementary Source of Information

There is an intellectual tension between the optimal experimental design point of view and the simple mean difference point of view for social experiments. The older optimal experimental design point of view stresses explicit models and the use of experiments to recover parameters of behavioral or "structural" models. The simple randomization point of view seeks to bypass models and produces – under certain conditions – a clean answer to a central question (*Q–1*): does the program work for participants? The two points of view can be reconciled if one is agnostic about the prior information at the disposal of analysts to design experiments. (Savage, et.al 1962). However, the benefits of randomization are less apparent when the goal is to recover trainee participation and continuation functions rather than if it is to recover the distribution of program outcome measures.

The potential conflict between the objectives of experimentation as a means of obtaining better estimates of a behavioral model and experimentation as a method for producing simple estimators of mean program impacts comes out forcefully when we consider using data from randomized experiments to estimate a behavioral model. To focus on main points, consider a program with two stages. $D_1 = 1$ if a person completes stage one; $= 0$ otherwise. $D_2 = 1$ if a person completes stage two; $= 0$ otherwise. Suppose that outcome Y can be written in the following form:

$$Y = \theta_0 + \theta_1 D_1 + \theta_2 D_1 D_2 + U. \tag{14}$$

The statistical problem is that D_1 and D_2 are stochastically dependent on U because unobservables in the outcome equation help determine D_1 and D_2. Randomizing at stage one makes D_1 independent of U. It does not guarantee that $D_1 D_2$ is stochastically independent of U.

The simple mean–difference estimator comparing outcomes of stage one

completers with outcomes of those randomized out, estimates, in large samples,

$$E(Y|D_1 = 1) - E(Y|D_1 = 0)$$
$$= \theta_1 + \theta_2 E(D_2|D_1 = 1).$$

In order to estimate θ_2 or θ_1, to estimate marginal effects of program completion at each stage, it is necessary to find an instrumental variable for D_1D_2.

Randomization on one coordinate only eliminates the need for one instrument to achieve this task. The appropriate stage at which the randomization should be implemented is an open question. The tradeoff between randomization as a source of instrumental variables and better nonexperimental sample design remains to be investigated. The optimal design of an experiment to estimate the parameters of equations like (14), or their extensions, in general would not entail simple randomization at one stage. The data generated as a byproduct of a one–shot randomization are only ideal for the estimation of models like (14) in the limited sense of requiring one fewer instrumental variable to consistently estimate θ_1 or θ_2 although this is a real benefit.

7. Summary

This paper critically examines the recent case for randomized social experimentation as a method for evaluating social programs. The method produces convincing answers to certain policy questions under certain assumptions about the behavior of agents and the questions of interest to program evaluators.

The method is ideal for evaluating social programs if attention focuses on estimating the mean effect of treatment on outcomes of the treated and one of the following set of assumptions holds:

AS–1: There is no effect of randomization on participation decisions

or

AS–2: If there is an effect of randomization on participation decisions, either (a) the effect of treatment is the same for all participants or (b) if agents differ in their response to treatments, their idiosyncratic responses to treatment do not influence their participation decisions.

If attention focuses on other features of social programs such as the determinants of participation, rejection or continuation decisions, randomized data possess no comparative advantage over stratified nonrandomized data. Even if AS–1 is true, experimental data cannot be used to investigate the distribution of program outcomes or its median without invoking additional "statistical" or "econometric" assumptions. In a multistage program randomized experimental data produce a "clean" (mean–difference) estimator of program impact only for outcomes defined conditional on the stage(s) where randomization is implemented. Statistical methods with their accompanying assumptions must still be used to evaluate outcomes at other stages and marginal outcomes for each stage.

Under assumptions that ensure it produces valid answers, the randomized experimental method bypasses the need to specify elaborate behavioral models. However, this makes experimental evidence an inflexible vehicle for predicting outcomes in environments different from those used to conduct the experiment. Interpolation and extrapolation replace model–based forecasting. However, such curve–fitting procedures may produce more convincing forecasts than ones produced from a controversial behavioral model.

Assumption AS–1 is not controversial in the context of randomized agricultural experimentation. This was the setting in which the Fisher model of experiments (1935) was developed. This model is the intellectual foundation for the recent case for

social experiments. AS–1 is more controversial in the context of randomized social experiments and is controversial even in the context of randomized clinical trials in medicine. Human agents may respond to randomization. These responses potentially threaten the reliability of experimental evidence. The evidence in Section 4 calls into question the validity of AS–1.

If that assumption is not valid, and if program participants respond differently to common treatments and these differences at least partly determine program participation decisions (so AS–2 is false), experimental methods do not even estimate the mean effect of treatment on the treated. In this case, randomized experimental methods answer the wrong question unless randomization is a permanent feature of the social program being evaluated. Data from randomized experiments cannot be used to estimate program participation, enrollment and continuation equations for ongoing programs.

References

Arnott, Richard and Joseph Stiglitz, "Randomization With Asymmetric Information", Rand Journal of Economics, Vol. 19, #3, Autumn 1988, pp. 344–362.

Ashenfelter, Orley, "Determining Participation in Income–Tested Programs", Journal of The American Statistical Association, September, 1983, Vol. 78, #383, pp. 517–525.

Ashenfelter, Orley and David Card, "Using The Longitudinal Structure of Earnings to Estimate The Effect of Training Programs", Review of Economics and Statistics, Vol. 67, 1985, pp. 648–660.

Cain, Glenn, "Regression and Selection Models To Improve Nonexperimental Comparisons", in C.A. Bennett and A.A. Lumsdaine, Evaluation and Experiment, New York, Dedhem Press, 1975.

Cain, Glenn and Harold Watts, "Summary and Overview", Final Chapter is in G. Cain and H. Watts, eds., Income Maintenance and Labor Supply, Chicago.

Cain, Glenn and Robert Wissoker, "A Reanalysis of Marital Stability In The Seattle–Denver Income Maintenance Experiment", American Journal of Sociology, Vol. 95, No. 5, March, 1990, pp. 1235–1269.

Conlisk, John, "Choice of Response Functional Form in Designing Subsidy Experiments", Econometrica, Vol. 41, 1973, pp. 643–656.

Conlisk, John and Harold Watts, "A Model For Optimizing Experimental Designs For Estimating Response Surfaces", American Statistical Association Proceedings, Social Statistics Section, 1969, pp. 150–156.

Cook, Thomas and Donald Campbell, Quasi–Experimentation: Design and Analysis Issues For Field Settings, Rand McNally Publishing Company, Chicago, 1979.

Cox, David R., Planning of Experiments, Wiley, New York, 1958.

Fisher, Ronald A., The Design of Experiments, Oliver and Boyd, London (1^{st} edition, 1935; 6th edition 1951).

Fraker, Thomas and Rebecca Maynard, "Evaluating Comparison Group Designs With Employment–Related Programs", Journal of Human Resources, Vol. 22, 1987, pp. 194–227.

Haavelmo, Trygve, "The Probability Approach in Econometrics", Econometrica, Vol. 12, Supplement, July 1944.

Ham, John, and Robert LaLonde, "Estimating The Effect of Training on The Incidence and Duration of Unemployment: Evidence on Disadvantaged Women From Experimental Data", University of Chicago, Graduate School of Business, 1989.

Hannan, Michael and Nancy B. Tuma, "A Reassessment Of The Effect of Income On Marital Dissolution in the Seattle–Denver Experiment", American Journal of Sociology, Vol. 95, No. 5, pp. 1270–1298.

Hausman, Jerry and David Wise, Social Experimentation, University of Chicago For NBER, 1985.

_____, "Technical Problems in Social Experimentation: Cost Versus Ease of Analysis", pp. 187–220 in J. Hausman and D. Wise, eds., Social Experimentation, University of Chicago Press For NBER.

Heckman, James J., "Dummy Endogenous Variables In A Simultaneous Equations Systems", Econometrica, Vol. 46, No. 3, pp. 931–961, July, 1978.

_____ (1990a), "Varieties of Selection Bias", American Economic Review, May, 1990.

_____ (1990b), "Alternative Approaches To The Evaluation of Social Programs: Econometric and Experimental Methods", Barcelona Lecture, World Congress of The Econometric Society, August 1990.

Heckman, James J. and Richard R. Robb, "Alternative Methods for Evaluating The Impact of Interventions: An Overview", Journal of Econometrics, Vol. 30, 1986, pp. 238–269.

_____, "Alternative Methods For Evaluating The Impact of Interventions", in Longitudinal Analysis of Labor Market Data, eds. J. Heckman and B. Singer, New York: Cambridge University Press, 185, pp. 156–245.

Heckman, James and V. J. Hotz, "Choosing Among Alternative Nonexperimental Methods For Estimating The Impact of Social Programs: The Case of Manpower Training", Journal of The American Statistical Association, Vol. 84, #408, December, 1989, pp. 862–880.

Heckman, James and Bo Honore', "The Empirical Content of The Roy Model", Econometrica, Vol. 58, September 1990, pp. 849–891.

Holland, Paul W., "Statistics and Causal Inference", Journal of The American Statistical Association, Vol. 81, No. 396, December, 1986, pp. 945–960.

Kramer, Michael S. and Stanley Shapiro, "Scientific Challenges in The Application of Randomized Trials", Journal of The American Medical Association, Nov. 16, 1984, Vol. 252, pp. 2739–2745.

LaLonde, Robert, "Evaluating The Econometric Evaluations of Training Programs with Experimental Data", American Economic Review, 1986, Vol. 76, pp. 604–620.

Lucas, Robert E., "Econometric Policy Evaluation: A Critique", Chapter 6 in Studies in Business Cycle Theory, MIT Press, 1981.

Manski, Charles, "The Selection Problem", forthcoming in C. Sims, editor, Advances in Econometrics: Proceedings of the Sixth World Congress of Econometrics, Cambridge University Press, 1991.

Marschak, Jacob, "Economic Measurements For Policy and Prediction", in Studies in Econometric Method, ed. by Wm. C. Hood and T.C. Koopman, Cowles Commission Monograph 13, New York: Wiley, 1953.

Orcutt, Guy and Orcutt, Alice, "Experiments For Income Maintenance Policies", American Economic Review, Vol. 58, No. 4, September, 1968, pp. 754–772.

Palaca, J., "AIDS Drug Trials Enter New Age", Science Magazine, October 6, 1989, pp. 19–21.

Rubin, Donald B., "Bayesian Inference For Causal Effects: The Role of Randomization", Annals of Statistics, Vol. 6, No.1, 1978, pp. 34–58.

Savage, L. J., et.al., The Foundations of Statistical Inference: A Symposium, M.S. Bartlett, editor, Wiley, New York, 1962, pp. 33–34.

Silvey, S. D., Optimal Design, Chapman Hall, London, 1980.

Tinbergen, J., Economic Policy: Principles and Design, North Holland, Amsterdam, 1956.