NBER TECHNICAL WORKING PAPER SERIES

RANDOMIZATION AND SOCIAL POLICY EVALUATION REVISITED

James J. Heckman

Technical Working Paper 107
http://www.nber.org/papers/t0107

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 1991, Revised January 2020

I would like to thank Rafeh Qureshi for comments made on this draft. This research was supported in part by: NIH grants NICHD R37HD065072 and NICHD R01HD054702; and the American Bar Foundation. The views expressed in this paper are solely those of the authors and do not necessarily represent those of the funders nor the official views of the National Institutes of Health. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Randomization and Social Policy Evaluation Revisited
James J. Heckman
NBER Technical Working Paper No. 107
July 1991, Revised January 2020
JEL No.

ABSTRACT

This paper examines the case for randomized controlled trials in economics. I revisit my previous paper "Randomization and Social Policy Evaluation" and update its message. I present a brief summary of the history of randomization in economics. I identify two waves of enthusiasm for the method as "Two Awakenings" because of the near-religious zeal associated with each wave. The First Wave substantially contributed to the development of microeconometrics because of the flawed nature of the experimental evidence. The Second Wave has improved experimental designs to avoid some of the technical statistical issues identified by econometricians in the wake of the First Wave. However, the deep conceptual issues about parameters estimated, and the economic interpretation and the policy relevance of the experimental results have not been addressed in the Second Wave.

James J. Heckman
Department of Economics
The University of Chicago
1126 E. 59th Street
Chicago, IL 60637
and IZA
and also NBER
jjh@uchicago.edu

Preamble

This paper updates my published 1992 paper, "Randomization and Social Policy Evaluation[1]" and places it in the context of the research that followed. The paper is still relevant for understanding the fundamental nature of experiments and what can be learned from even "ideal" experiments with no attrition, non-response, and stratification on the outcome variables of interest. It is worth revisiting in light of the continuing controversies surrounding the role of randomization in development economics. The conceptual points made here have not been addressed in the literature, even though many issues of implementation have.

This preface provides some perspective on the history of field experiments and the origins of the experimental movement in economics. The history of field experimentation in economics since 1965 can be classified into two eras: (1) The early wave that used experiments to settle important policy debates where nonexperimental evidence was ambiguous; and (2) the revival of experimentation in development economics that culminated in the 2019 Nobel Prize in Economics. Each era has been marked by a near-religious zeal for the methodology of Randomized Control Trials (RCTs). Accordingly, I name both the eras, "Great Awakenings," in honor of two religious revivals that shaped Protestant churches in North America in the 18th and 19th centuries, and in recognition of the zeal for methodological purity in both eras in economics.

The First Great Awakening arose in the push to evaluate the manpower, education, and health programs launched by Lyndon Johnson's War on Poverty. The Second Great Awakening came in development economics in the wake of a variety of micro programs targeted to less-developed countries funded by influential NGOs, billionaires, and various international institutions. Few of the hard lessons learned about the limitations of social experiments from the First Great Awakening are acknowledged by the economists promoting the Second Awakening. The career incentives of the new generation argue against examining and citing the contributions and lessons of the First Awakening, which ended in substantial qualifi-

[1]See Heckman (1992).

cation of the alleged claim of "transparent results" and eventual decline in the uncritical enthusiasm for RCTs. The Second Awakening will likely suffer the same fate.

The First Awakening

Long before randomization became *de rigueur* in the field of development in the First Great Awakening, it was advocated for evaluating a variety of social programs, educational interventions, workforce training programs, and welfare reforms.

In the First Wave, leading evaluation firms, such as Westat, Mathematica, SRI, Abt Associates, and MDRC, addressed the mandate of the office of Economic Opportunity (OEO) that administered Lyndon Johnson's War on Poverty to evaluate a raft of newly launched social programs. The emphasis on evaluation percolated across many U.S. federal agencies.

This early thrust for evaluation led to the collection of novel panel micro data sets that continue to guide understanding of society and are now widely emulated around the world. The First Awakening also fostered new methodologies to analyze the serious problems that plagued the experiments conducted in the First Wave.

The first wide-scale use of randomization in economics was in evaluating Negative Income Tax (NIT) programs. These programs were proposed by Milton Friedman[2] and others as an alternative to the cumbersome welfare transfer programs of the day that heavily taxed low income workers by substantially reducing benefits for each dollar earned. The NIT was designed to replace the patchwork welfare system of the 1960s by giving a lump sum transfer to the poor and taxing additional earnings at a uniform low rate over the whole income schedule. The policy question was whether imposition of NIT would substantially reduce labor supply. The answer depended on the relative strength of income and substitution effects. Transfers would reduce labor supply through an income effect. The lowered tax rate on earnings would encourage it through a substitution effect. The existing nonexperimental estimates of the income and substitution effects ranged all over the place, as documented in

[2]See Friedman (2009), reissued.

the introductory chapter of Cain and Watts (1973).

In the early 1960s, Heather Ross, then a graduate student at MIT, proposed a large scale randomized trial to gauge the effects of NIT. The Office of Economic Opportunity accepted her proposal and funded it. Many economic consulting firms rose to the challenge. The first NIT experiment was launched in 1968.

The early researchers waded into deep waters and sometimes got in over their heads. The initial designs were flawed. Selection bias riddled the study. Attrition and noncompliance was high. Ironically, analyzing NIT data helped to launch the then nascent field of microeconometrics. The era culminated in John Cogan's testimony before the U.S. Congress[3], in which he reanalyzed the data from the NIT experiment using the newly-developed techniques of microeconometrics. These methods were later recognized by the Nobel Prize Committee in 2000.

Cogan's testimony challenged the "transparent" evidence from the experiment pointing out a variety of selection biases. He showed negative impacts on labor supply that were substantially larger than the trivial impacts found from the "transparent" experimental comparisons of the mean differences between treatments and controls. At those hearings, Senator Daniel Patrick Moynihan expressed dismay over the low quality of the "transparent" experimental evidence, as revealed by Cogan's analysis, and gratitude to Cogan for presenting an honest report of what the experiment actually demonstrated using nonexperimental methods to analyze the flawed experimental data.

The Second Awakening

The Second Wave is in its zenith. The enthusiasm for experimentation has led NGOs, foundations, and governments to mandate its application. Whereas the First Wave was motivated by the desire to address major social questions, the Second Wave has a more methodological focus. It is part and parcel of a professional obsession in the field of economics to obtain

[3]Congress (1978)

"causal effects," even if the effects being identified are without social significance and/or economic meaning.[4] Miniaturist studies became praised as the ideal for rigorous empirical economics. Asking and trying to answer big and important questions was discredited in pursuit of clean answers to small questions of little policy consequence. Indeed, the Nobel Prize Committee in 2019 lauded practitioners in the Second Wave for focusing on "smaller, more manageable problems.[5]" The award was for methodological purity and "manageability" rather than for substance.

It is useful to cast the quest of many applied economists marching in the parade of the Second Wave in terms of a traditional regression framework. Let Y be an outcome of interest. Suppose

$$Y = X\beta + D\alpha + U$$

where X is a vector of observed control variables, D is an indicator if treatment is received ($D = 1$ if treated, $D = 0$ if not), and U is correlated with D. α is "the effect" of treatment controlling for X and U. If we fail to control for X and U, correlational estimates of α are biased with the sign of the bias determined by the sign of the correlation between U and D controlling for X. Randomization avoids this bias if it is properly conducted.

As in the recent instrumental variables literature, in the Second Awakening eliminating this bias is the paramount issue usually to the exclusion of asking whether α answers any important question – either in theory or practice. In the First Awakening, that issue was front and center.

The revised paper presented here, and a follow-up paper by Heckman and Smith (1995), were written after the First Wave of enthusiasm for RCTs and before the Second Wave. Both papers are relevant today. The fact that the Second Wave emerged is a tribute either to the bad writing of those papers, or to the demonstrated ability of economists to ignore

[4]See the essay by Deaton in this volume.
[5]Royal Academy of Sciences (2019).

hard-won lessons from the past and the strong career incentives to pour old wine into new bottles and forget its sources. I now turn to the original paper.

1 Introduction

This paper considers the benefits and limitations of *randomized* social experimentation as a tool for evaluating social programs.[6] 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; that is, 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 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 policymaking. The goal of social experimentation, as envisioned by Conlisk and Watts (1969) and Conlisk (1973), was to develop better estimates of the structural equations needed to

[6]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 (1985a), and the other chapters in this volume for statements on problems of attrition, spillover effects, and so forth.

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 individuals 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). Extrapolating 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, 1985b).

The recent case for randomized social experiments represent a dramatic retreat from the ambitious program of "optimal" social policy analysis that was never fully embraced by most economists and was not embraced at all by other social scientists. Considerable skepticism had 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 stimulate social programs not yet in place. Influential studies by LaLonde (1986) and Fraker and Maynard (1987) convinced many that econometric and statistical methods are incapable of estimating true program impacts from nonrandomized 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 (whether 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 evaluations.[7] 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, that is, 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 real possibility. And there is evidence that it is an important phenomenon.

In addition, advocates of randomization implicitly assume that certain mean differences in outcomes are invariably the objects of interest in performing an evaluation. In fact, there are many parameters of potential interest, only some of which can be cast into a mean-difference framework. Experimental methods *cannot* estimate median differences or other "quantile treatment effects" without invoking stronger assumptions than are required to recover means. 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. Many

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

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 (that is, 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 rarely been implemented, probably because it would drastically alter the program being evaluated.[8] But if only one randomization can be conducted, an evaluation of all stages of a multistage program entails the use of the very controversial econometric methodology sought to be avoided in the recent case for social experimentation.

The purpose of this paper is to clarify arguments for and against randomized social experiments. In order to focus the discussion, I first present a prototypical social program and consider what features of the program are of interest to policy evaluators. In the second section, I discuss the difficulties that arise in determining program features of interest. A precise statement of the evaluation problem is given. In the following section, I state the case for simple randomization; then I consider the implicit behavioral assumptions that underlie the case and the conditions under which they hold. I also discuss what can and cannot be learned from a randomized social experiment even under ideal conditions. In the fourth section I present some indirect evidence on the validity of the assumptions for the case of a recent evaluation of the Job Training Partnership Act (JTPA). I also consider some parallel studies of their validity in randomized clinical trials literature in medicine. In the fifth section I discuss the issue of choosing the appropriate stage at which one should randomize in a multistage program. In the sixth section I discuss the tension between the new and the old cases for social experimentation. The final section summarizes the argument.

[8]See, however, the evaluation of the ABC program: Ramey et al. (1976), which has multistage randomization.

2 Questions of Interest in Evaluating a Prototypical Social Program

The prototype considered here is a manpower training program similar to the JTPA program described by Heckman, Ichimura, Smith, and Todd (1998). That prototypical program offered a menu of training options to potential trainees. Specific job-related skills may be learned as well as general skills (such as reading, writing, 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 an activity. Some specific skill programs entail working for an employer at a subsidized wage (that is, on-the-job-training).

Individuals who receive training proceed through the following steps: they (1) apply; (2) are accepted; (3) are placed in specific training sequence; (4) are reviewed; (5) are certified in a skill; and (6) are placed with 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 were paid by the U.S. government on the basis of the quality of the placement of their trainees. Quality was measured in part by the wages received over a specified period of time after trainees complete their training program (for example, six months). Managers thus had an incentive to train persons who are likely to attain high-quality placement and who can achieve the status at low cost to the center. Trainees received compensation (subsidies) while in the program. Training centers recruited 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 the 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. Other questions of interest concern 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?

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?

Q-5 What are the effects of family background, center profit rates, subsidies, and local labor market conditions on the decision to drop out from a program and the length of time taken to complete the program?

Q-6 What are the effects of labor market conditions, subsidies, profit rates, and so forth on placement rates and wage and hour levels attained at placement?

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 policymakers. The central evaluation problem is how to obtain convincing answers to them.

3 The Evaluation Problem

To characterize the essential features of the evaluation of the evaluation problem, it is helpful to concentrate on only on a few of the questions listed above. I focus attention on questions Q-1 and Q-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 that there is 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.[9]

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 that are not germane here. Structural 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 in turn 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

[9]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 Heckman et al., 1998).

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, for example, 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 variations in wage and income levels across subjects so that precise estimates of wage and income effects could be obtained. Cain and Watts (1973) argued that in cross-section data, variation in wages and income 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 response 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 are able to be measured. Both the 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 these programs that might affect programs participation.

3.1 A Model of Program Evaluation

To be more specific, 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:

$$\begin{aligned} Y &= Y_1 \text{ if } D = 1 \\ Y &= Y_0 \text{ if } D = 0 \end{aligned} \tag{1}$$

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 reference" by some statisticians (see, for example, Holland, 1986). Let Y_1 and Y_0 be determined by X_1 and X_0 respectively. Presumably X_1 induces 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:

$$Y_1 = g_1(X_1), \tag{2a}$$

$$Y_0 = g_0(X_0). \tag{2b}$$

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

$$Y_1 = X_1\beta_1 \tag{2a$'$}$$

and

$$Y_0 = X_0\beta_0 \tag{2b$'$}$$

respectively.

Let Z be variables determining program participation. If

$$Z \in \Psi, D = 1; \quad Z \notin \Psi, D = 0, \tag{3}$$

where Ψ is a set of 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. 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′ (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). 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 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 answer fully 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 outcomes 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.

Let subscript a denote available information. Thus, C_a 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 $C_a = c_a$ is

$$F(y_0, y_1, d \mid c_a) = \Pr(Y_0 \leq y_0, Y_1 \leq y_1, D = d \mid C_a = c_a), \tag{4}$$

where I follow convention by denoting random variables by uppercase letters and their realization by lowercase letters. If (4) can be determined, and the distribution of C_a is known, it is possible to answer questions Q-1, Q-2 and Q-3′ in the following sense: one can determine the population distribution of Y_0, Y_1 and the population *distribution* of the gross gain from the program participation,

$$\Delta = Y_1 - Y_0,$$

and one can write out the probability of the event $D = d$ given Z_a.

3.2 The Parameters of Interest in Program Evaluation

We can answer Q-1 if we can identify

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

and hence

$$F(\delta \mid D = 1, c_a)$$

(the distribution of the effect of treatment on the treated, where δ is the lowercase version of Δ). One can answer Q-2 if we know

$$F(y_0, y_1 \mid c_a), \tag{5}$$

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, $F(y_0 \mid c_a)$ to another $F(y_1 \mid c_a)$. The answer to Q-3′

can be achieved by computing from (4) the probability of participation:

$$\Pr(D = 1 \mid c_a) = F(d \mid c_a).$$

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

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

and the answer to Q-2 as

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

although in principle knowledge of the full distribution of Δ, or some other features besides the mean (for example, the median), might be desirable.

Even if the means in (6) and (7) were zero, it is of interest to know what fraction of participants or of the population would benefit from a program. This would require knowledge of $F(\delta \mid D = 1, c_a)$ or $F(\delta \mid c_a)$, respectively. In order to ascertain the existence of "cream skimming" (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

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

or

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

other than the means of Y_1 and Y_0. To answer many questions, knowledge of mean differences in 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 posed in the recent literature do not produce data sufficient for this task.

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

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

the distribution of outcomes for participants, and

$$F(y_0 \mid D = 0, c_a),$$

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

$$\Pr(D = 1 \mid c_a) = F(d \mid c_a),$$

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 $C_a = c_a$

$$F(y_0 \mid D = 1, c_a) \neq F(y_0 \mid D = 0, c_a) \tag{8a}$$

and

$$F(y_1 \mid D = 1, c_a) \neq F(y_1 \mid D = 0, c_a), \tag{8b}$$

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:

$$E(\Delta \mid D = 1, c_a) \neq E(Y_1 \mid D = 1, c_a) - E(Y_0 \mid D = 0, c_a) \tag{9a}$$

$$E(\Delta \mid c_a) \neq E(Y_1 \mid c_a) - E(Y_0 \mid c_a), \tag{9b}$$

that is, 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 nonparticipants even after adjusting for C_a.

Many methods have been proposed for solving the selection problem either for means or for entire distributions. Heckman and Honoré (1990), Heckman and Robb (1986), Heckman and Robb (1985), Heckman (1990b), Heckman (1990a), and Heckman, Smith, and Clements (1997) offer alternative comprehensive treatments of the various approaches to this problem in econometrics and statistics. Some untestable a priori assumptions must be invoked to recover the missing components of the distribution. Constructing these counterfactuals inevitably generates controversy.

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 nonexperimental 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 standard model specification tests for their nonexperimental alternative estimates. When such tests are performed, they estimate all but the nonexperimental models that reproduce the inference obtained by experimental methods.

There is, nonetheless, a kernel of truth in the criticism of LaLonde (1986) and Fraker and Maynard (1987). Each test of a nonexperimental model proposed by Heckman and Hotz (1989) has limitations. Test of overidentifying features of a model can be rendered worthless by changing the model to a just-identified form, a criticism that also arises in application of the Durbin-Wu-Hausman test.[10]

All nonexperimental methods are based on some maintained, untestable assumption.

[10]See Durbin (1954); Hausman (1978); Wu (1973).

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 responses of agents to randomization. These assumptions are different from but not necessarily more credible than the assumptions maintained in the nonexperimental econometrics and statistics literatures.

4 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 question Q-1 and Q-2—the "causal problem" as defined by statisticians. (See Cox, 1958; Fisher, 1935; Holland, 1986; Rubin, 1978). 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 nonexperimental regime. Let D^* denote a variable indicating if an agent is at risk for ran-

domization (that is, if the agent applied and was accepted in a regime of random selection):

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

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

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

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 United States where full disclosure of programs operating rules is required by law. Even if it were possible to surprise potential trainees, it would not be possible 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 ($p = \Pr(A = 1)$), that is, it includes the effect of randomization on agent and center choices.

Proponents of randomization invoke the assumption that

$$\Pr(D = 1 \mid c) = \Pr(D^* = 1 \mid c, p), \tag{AS-1}$$

or assume that it is "practically" true.[11]

[11] Failure of this assumption is an instance of the Marschak (1953) – Lucas (1981) Critique applied to social experimentation. It is also an instance of a "Hawthorne" effect. See Cook and Campbell (1979).

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,[12] 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 also may 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 developed. However, the Fisher model is potentially a misleading paradigm for social science. Humans act purposively, and their behavior is likely to be altered by introducing randomization in 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,

$$F(y_1, c \mid A = 1) = F(y_1, c \mid D^* = 1) = F(y_1, c \mid D = 1), \tag{10a}$$

$$F(y_0, c \mid A = 0) = F(y_0, c \mid D^* = 1) = F(y_0, c \mid D = 1), \tag{10b}$$

[12]This can arise even if agents are risk averse by convexifying a noncovex problem. See Arnott and Stiglitz (1988).

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

Simple mean difference estimators between participants and randomized-out nonparticipants 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,

$$Y_1 = \alpha + Y_0. \tag{12}$$

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

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

The dummy endogenous variable model is widely used in applied work. Reliance on this model strengthens the popular case for randomization. Q-1 and Q-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):

$$Y_1 = Y_0 + (\alpha + \Xi), \tag{13a}$$

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

$$E(\Xi \mid D) = 0, \tag{13b}$$

then (11) remains true. If potential trainees and training centers do not know the trainees' gain from the program in advance of their enrollment in the program, and they use $\alpha + \Xi$ 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 Q-1.

It is important to note how limited are the data obtained from an "ideal" social experiment (that is, 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, unless one makes the extreme assumption of rank invariance, i.e., that the rank of persons in the Y_1 distribution are the same in the Y_0 distribution.[13] 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. Answer to most of the questions listed in the first section still require application of econometric procedures with their attendant controversial assumptions.

If assumption (AS-1) is not satisfied, the final equalities in (10a) and (10b) 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; that is,

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

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 agents' response to programs is heterogenous ($\Xi \neq 0$) and

[13] See Heckman, Smith, and Clements (1997).

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 noncontroversial 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.

5 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 one 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, although it is indirect. In response to the wide variability in estimates of the impact of manpower programs derived from nonexperimental estimators by LaLonde (1986) and Fraker and Maynard (1987), the U.S. Department of Labor financed a large-scale experimental evaluation of the large-scale Job Training Partnership Act (JTPA), which was the main vehicle for providing government training in the United States. Randomized evaluation was implemented in a variety of sites. The organization implementing this experiment—the Manpower Demonstration Research

Corporation (MDRC)—is an ardent and effective advocate for the use of randomization as a method for evaluating social programs.

A report by this organization (Doolittle and Traeger, 1990) gives some information from which it is possible to do a rough revealed preference analysis.[14] Job training in the United States in the late 1980s and early 1990s was organized through geographically decentralized centers. These centers received 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 percent to 10 percent of the total operating costs of the centers.

In attempting to enroll geographically dispersed sites, MDRC experienced a training center refusal rate in excess of 90 percent. 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 applicant pool, which 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 behavior ruled out by assumption (AS-1). In attempting to entice centers to participate, MDRC had to reduce the randomized rejection probability from $\frac{1}{2}$ to as low as $\frac{1}{6}$ 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 sevenfold in order to get any centers to participate in the experiment. the MDRC analysts concluded:

> *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*

[14] Hotz (1992) also summarizes their discussion

Table 1: Percentage of local JTPA agencies citing specific concerns about participating in the experiment

Concern	Percentage 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	54.5
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 local JTPA agencies contacted about possible participation in the National JTPA Study. From Doolittle and Traeger (1990). Copyright 1989, 1990 by the Manpower Demonstration Research Corporation and used with its permission.

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

> *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)

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"(p.123).

During the experiment conducted at Corpus Christi, Texas, center administrators successfully petitioned the government of Texas for a waiver of its performance standards on the

ground that the experiment disrupted center operations. 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 the JTPA system as a whole. At least the data can be used to provide a lower-bound estimate of the major impact of disruption.

Randomization is also controversial in clinical trials in medicine which are sometimes held up as a paragon for empirical social science.[15] 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* (1989), notes that AIDS patients denied potentially life-saving drugs took 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 to 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 nonexperimental studies. They discuss one study of drugs administered to children afflicted with a disease. The study had two components. The nonexperimental phase of the study had a 4 percent refusal rate, while 34 percent of a subsample of the same parents refused to participate in a randomized subtrial, although the treatments were equally nonthreatening.

These authors cite evidence suggesting that non-response to randomization is selective. In a study of treatment of adults of cirrhosis, no effect of the treatment was found for participants in a randomized trial. But the death rates for those randomized out of the

[15]See, for example, Ashenfelter and Card (1985).

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.

6 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 multiple randomization has rarely been performed likely indicates that it would exacerbate the problem of randomization bias discussed in the two previous sections. 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, nonexperimental "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, 1990) or controversial (Cain and Wissoker, 1990; Hannan and Tuma Brandon, 1990) analyses of randomized experimental data that have appeared in the recent literature.

Moreover, for some of the questions posed at the beginning of the paper, it is not obvious that randomization is the method of choice for securing convincing answers. Many of the

questions listed there 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 income at the extremes of the eligible population.[16]

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 and these conditions would be desirable. 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 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 center from bearing the political cost 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, it still encounters the problem that people self select. Assuming that eligibility does not disrupt the fundamental program parameters, the simple mean-difference parameter comparing the eligible with the ineligible identifies $E(Y_1 - Y_0|D = 1)P$, where P is the probability of participation in the program through voluntary selection in the program in the absence of an experiment. Dividing by P, one can identify treatment on the treated.

[16]This remark assumes a linear model. For optimal designs in nonlinear models see, for example, Silvey (1980).

7 The Tension between the Case for Social Experiments as a Substitute for Behavioral Models and Social Experiments as 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 toward 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 one 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 (see Savage, 1962). However, the benefits of randomization are less apparent when the goal is to recover trainee participation and continuation functions 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. 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,

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

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 $D_1 D_2$.

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 trade-off between randomization as an instrumental variable and better nonexperimental sample design remains to be investigated. The optimal design of an experiment to estimate the parameters of (14) in general would not entail simple randomization at one stage. The data generated as a by-product of a one-shot randomization are only ideal for the estimation of models like (14) in the limited sense of requiring one less instrumental variable to consistently estimate θ_1 or θ_2, although this is a real benefit.

8 Summary of the 1992 Paper

This paper critically examines the case made in the First Awakening for randomized social experimentation as a method for evaluating social programs. The method produces convincing answers to certain policy questions under strong 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 if 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 possesses 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 their 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 conditionally 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 that 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 (1935) model of experiments was developed. That model is the intellectual foundation for recent case for social experiments, although the recent literature in economics often misattributes it to statisticians of the 1970s. Assumption (AS-1) is more controversial even in the context of randomized clinical trials in medicine. Human agents may respond to randomization, and these responses potentially threaten the reliability of experimental evidence. The evidence on randomization bias

presented earlier calls into question the validity of (AS-1).

If that assumption is not valid, and if the program participants respond differently to common treatments and those differences at least partly determine program participation decisions (so that (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.

Post-Script, 2019

I stand by my discussion of the conceptual issues raised in this paper and my companion paper with Smith (1995).[17] The points made are all valid today and have largely been ignored in the recent "Second Awakening" revival in development economics. There are many papers written after these papers that establish or reiterate the points made here. In addition to failing to learn from the past, the Randomistas are ungenerous to the true pioneers of field experiments.

In subsequent work, Heckman and Smith (1998) develop the point that self-selection into a program generates information about agent *ex ante* perceptions of program benefits.[18] These subjective evaluations are arguably more important than the "objective" evaluations (δ) emphasized by statisticians who treat "non-compliance" as a problem rather than a source of information. This information would be suppressed if persons were forced to go into treatment or control status. This point is yet one more example of the benefits of using economics to design and evaluate social programs.

In later work, Heckman, Hohmann, Smith, and Khoo (2000) consider *substitution bias*

[17] I have since amplified these points in Heckman, Ichimura, and Todd (1997), Heckman, LaLonde, and Smith (1999), and Heckman and Vytlacil (2007).

[18] Thus, as noted by Heckman and Smith (1998), the pain and suffering of a medical trial may outweigh its benefits for survival.

as a major threat to straightforward interpretation of experiments. If agents have access to alternative programs, persons eligible to participate in a program and persons ineligible may choose to participate in an alternative program. The "transparent" mean difference between treatments and controls does not compare the effect of treatment with no treatment, but instead, the effect of treatment vs the best alternative which may in fact be better than the program being evaluated. Our (2000) paper documents the pervasiveness of the problem. Kline and Walters (2016) give a recent demonstration of the problem of substitution bias. The "transparent" mean difference estimator from a recent experimental evaluation of Head Start suggested that the program had no impact on disadvantaged children. A more careful analysis accounting for substitution bias using microeconometric methods shows a strong effect. Their paper echoes the 1978 finding of Cogan regarding the NIT.

Banerjee and Duflo (2009) respond to the points raised in my 1992 paper, as do Athey and Imbens (2017). They claim that its criticisms no longer apply due to improved survey design and implementation methodology. However, they do not discuss many basic interpretive or conceptual points in my 1992 paper or its 1995 companion, or the inability of experimental mean difference comparisons to answer the range of policy-relevant treatment effects discussed in my papers and in subsequent research (Heckman, 2008).

The literature after my 1992 paper has produced considerable evidence on the inadequacy of experimental evidence in many fields. Sanson-Fisher, Bonevski, Green, and D'Este (2007) show that experiments are fundamentally too limited in scope to consider impact evaluations, such as women's empowerment. Concato and Horwitz (2018) survey the consensus in medicine.[19] It has switched away from reliance on RCTs as the "gold standard," which they say was the party line in the 1990s in medicine. They present many papers discussing limitations of randomized experiments in medicine: (Horwitz, 1996; Feinstein and Horwitz, 1997; Concato and Horwitz, 2004; Concato, 2012, 2013; Horwitz and Singer, 2017; Shahar, 1997; Sehon and Stanley, 2003; Chakravarty and Fries, 2006; Worrall, 2007; Rawlins,

[19]The "paragon" cited by Ashenfelter and Card (1985).

2008; Borgerson, 2009; Frieden, 2017). Czibor, Jimenez-Gomez, and List (2019) is a recent cautionary paper for experimental economists that reiterates the points of my 1992 paper. It highlights serious problems in experimental economics and what devout experimentalists need to be wary of.

The causal models advocated in the recent program evaluation literature are motivated by the experiment as an ideal. They do not clearly specify the theoretical mechanisms determining the sets of possible counterfactual outcomes, how hypothetical counterfactuals are realized or how hypothetical interventions are implemented except to compare "randomized" with "nonrandomized" interventions. They focus on outcomes, leaving the model for selecting outcomes and the preferences of agents over expected outcomes unspecified.[20]

Those who ignore intellectual history are condemned to repeat past mistakes. The Second Wave will pass as economists relearn the lessons of the past.

[20]See Heckman (2008).

References

Arnott, R. and J. E. Stiglitz (1988, Autumn). Randomization with asymmetric information. *RAND Journal of Economics 19*(3), 344–362.

Ashenfelter, O. C. (1983, September). Determining participation in income-tested social programs. *Journal of the American Statistical Association 78*(383), 517–525.

Ashenfelter, O. C. and D. Card (1985, November). Using the longitudinal structure of earnings to estimate the effect of training programs. *Review of Economics and Statistics 67*(4), 648–660.

Athey, S. and G. W. Imbens (2017). The econometrics of randomized experiments. Volume 1 of *Handbook of Economic Field Experiments*, Chapter 3, pp. 73–140. North-Holland.

Banerjee, A. V. and E. Duflo (2009, April). The experimental approach to development economics. *Annual Review of Economics 1*, 151–178.

Borgerson, K. (2009). Valuing evidence: Bias and the evidence hierarchy of evidence-based medicine. *Perspectives in Biology and Medicine 52*(2), 218–233.

Cain, G. G. (1975). Regression and selection models to improve nonexperimental comparisons. In C. Bennett and A. A. Lumsdaine (Eds.), *Evaluation and Experiment : Some Critical Issues in Assessing Social Programs*, pp. 297–317. New York: Academic Press.

Cain, G. G. and H. W. Watts (1973). Summary and overview. In *Income Maintenance and Labor Supply: Econometric Studies*. Chicago: Rand McNally College Publishing Company.

Cain, G. G. and D. A. Wissoker (1990, March). A reanalyis of marital stability in the Seattle-Denver income-maintenance experiment. *American Journal of Sociology 95*(5), 1235–1269.

Chakravarty, E. F. and J. F. Fries (2006). Science as experiment; science as observation. *Nature Clinical Practice Rheumatology 2*(6), 286.

Concato, J. (2012). Is it time for medicine-based evidence? *The Journal of the American Medical Association 307*(15), 1641–1643.

Concato, J. (2013). Study design and "evidence" in patient-oriented research. *American Journal of Respiratory and Critical Care Medicine 187*(11), 1167–1172.

Concato, J. and R. I. Horwitz (2004). Beyond randomised versus observational studies. *The Lancet 363*(9422), 1660–1661.

Concato, J. and R. I. Horwitz (2018). Randomized trials and evidence in medicine: A commentary on Deaton and Cartwright. *Social Science & Medicine 210*, 32–36.

Congress, United States, S. C. o. F. S. o. P. A. (1978). *Welfare research and experimentation: Hearings before the Subcommittee on Public Assistance of the Committee on Finance, United States Senate, Ninety-fifth Congress, second session, November 15, 16, and 17*. Washington: U.S. Government.

Conlisk, J. (1973, July). Choice of response functional form in designing subsidy experiments. *Econometrica 41*(4), 643–656.

Conlisk, J. and H. Watts (1969, August). A model for optimizing experimental designs for estimating response surfaces. *American Statistical Association Proceedings, Social Statistics Section*, 150–156.

Cook, T. D. and D. T. Campbell (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Chicago: Rand McNally College Publishing Company.

Cox, D. R. (1958). *Planning of Experiments*. New York: Wiley.

Czibor, E., D. Jimenez-Gomez, and J. A. List (2019). The dozen things experimental economists should do (more of). NBER Working Paper 25451.

Doolittle, F. C. and L. Traeger (1990). *Implementing the National JTPA Study*. New York: Manpower Demonstration Research Corporation.

Durbin, J. (1954). Errors in variables. *Review of the International Statistical Institute 22*, 23–32.

Feinstein, A. R. and R. I. Horwitz (1997). Problems in the "evidence" of "evidence-based medicine". *The American Journal of Medicine 103*(6), 529 –535.

Fisher, R. A. (1935). *The Design of Experiments*. London: Oliver and Boyd.

Fraker, T. and R. Maynard (1987, Spring). The adequacy of comparison group designs for evaluations of employment-related programs. *Journal of Human Resources 22*(2), 194–227.

Frieden, T. R. (2017). Evidence for health decision making — beyond randomized, controlled trials. *New England Journal of Medicine 377*, 465–475.

Friedman, M. (2009). *Capitalism and Freedom*. University of Chicago Press.

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica 12*(Supplement), iii–vi and 1–115.

Ham, J. C. and R. J. LaLonde (1990). Using social experiments to estimate the effect of training on transition rates. In J. Hartog, G. Ridder, and J. Theeuwes (Eds.), *Panel Data and Labor Market Studies*, pp. 157–172. Oxford, UK: North-Holland.

Hannan, M. T. and N. Tuma Brandon (1990, March). A reassessment of the effect of income maintenance on marital dissolution in the Seattle-Denver experiment. *American Journal of Sociology 95*(5), 1270–1298.

Hausman, J. A. (1978, November). Specification tests in econometrics. *Econometrica 46*(6), 1251–1272.

Hausman, J. A. and D. A. Wise (1985a). *Social Experimentation*. Chicago: University of Chicago Press.

Hausman, J. A. and D. A. Wise (1985b). Technical problems in social experimentation: Cost versus ease of analysis. In J. A. Hausman and D. A. Wise (Eds.), *Social Experimentation*, pp. 187–220. Chicago: University of Chicago Press.

Heckman, J. J. (1978, July). Dummy endogenous variables in a simultaneous equation system. *Econometrica 46*(4), 931–959.

Heckman, J. J. (1990a). Alternative approaches to the evaluation of social programs: Econometrics and experimental methods. Lecture, Sixth World Meetings of the Econometric Society, Barcelona, Spain.

Heckman, J. J. (1990b, May). Varieties of selection bias. *American Economic Review 80*(2: Papers and Proceedings of the Hundred and Second Annual Meeting of the American Economic Association), 313–318.

Heckman, J. J. (1992). Randomization and social policy evaluation. In C. F. Manski and I. Garfinkel (Eds.), *Evaluating Welfare and Training Programs*, Chapter 5, pp. 201–230. Cambridge, MA: Harvard University Press.

Heckman, J. J. (2008, April). Econometric causality. *International Statistical Review 76*(1), 1–27.

Heckman, J. J. and O. Ashenfelter (1973). Estimating labor supply functions. In G. G. Cain and H. Watts (Eds.), *Income Maintenance and Labor Supply*. Academic Press.

Heckman, J. J., N. Hohmann, J. Smith, and M. Khoo (2000, May). Substitution and dropout bias in social experiments: A study of an influential social experiment. *Quarterly Journal of Economics 115*(2), 651–694.

Heckman, J. J. and B. E. Honoré (1990, September). The empirical content of the Roy model. *Econometrica 58*(5), 1121–1149.

Heckman, J. J. and V. J. Hotz (1989, December). Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of Manpower Training. *Journal of the American Statistical Association 84*(408), 862–874. Rejoinder also published in Vol. 84, No. 408, (Dec. 1989).

Heckman, J. J., H. Ichimura, J. Smith, and P. E. Todd (1998, September). Characterizing selection bias using experimental data. *Econometrica 66*(5), 1017–1098.

Heckman, J. J., H. Ichimura, and P. E. Todd (1997, October). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies 64*(4), 605–654. Special Issue: Evaluation of Training and Other Social Programmes.

Heckman, J. J., R. J. LaLonde, and J. A. Smith (1999). The economics and econometrics of active labor market programs. In O. C. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*, Volume 3A, Chapter 31, pp. 1865–2097. New York: North-Holland.

Heckman, J. J., L. J. Lochner, and C. Taber (1998, January). Explaining rising wage inequality: Explorations with a dynamic general equilibrium model of labor earnings with heterogeneous agents. *Review of Economic Dynamics 1*(1), 1–58.

Heckman, J. J. and R. Robb (1985). Alternative methods for evaluating the impact of interventions. In J. J. Heckman and B. S. Singer (Eds.), *Longitudinal Analysis of Labor Market Data*, Volume 10, pp. 156–245. New York: Cambridge University Press.

Heckman, J. J. and R. Robb (1986). Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes. In H. Wainer (Ed.), *Drawing Inferences from Self-Selected Samples*, pp. 63–107. New York: Springer-Verlag. Reprinted in 2000, Mahwah, NJ: Lawrence Erlbaum Associates.

Heckman, J. J. and J. A. Smith (1995, Spring). Assessing the case for social experiments. *Journal of Economic Perspectives 9*(2), 85–110.

Heckman, J. J. and J. A. Smith (1998). Evaluating the welfare state. In S. Strom (Ed.), *Econometrics and Economic Theory in the Twentieth Century: The Ragnar Frisch Centennial Symposium*, pp. 241–318. New York: Cambridge University Press.

Heckman, J. J., J. A. Smith, and N. Clements (1997, October). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *Review of Economic Studies 64*(4), 487–535.

Heckman, J. J. and E. J. Vytlacil (2007). Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation. In J. J. Heckman and E. E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6B, Chapter 70, pp. 4779–4874. Amsterdam: Elsevier B. V.

Holland, P. W. (1986, December). Statistics and causal inference. *Journal of the American Statistical Association 81*(396), 945–960.

Horwitz, R. I. (1996). The dark side of evidence-based medicine. *Cleveland Clinic Journal of Medicine 63*(6), 320–323.

Horwitz, R. I. and B. H. Singer (2017). Why evidence-based medicine failed in patient care and medicine-based evidence will succeed. *Journal of Clinical Epidemiology 84*, 14–17.

Hotz, V. J. (1992). Designing an evaluation of the Job Training Partnership Act. In C. Manski and I. Garfinkel (Eds.), *Evaluating Welfare and Training Programs*, pp. 76–114. Cambridge, MA: Harvard University Press.

Kline, P. and C. Walters (2016). Evaluating public programs with close substitutes: The case of Head Start. Quarterly Journal of Economics *131*(4), 1795–1848.

Kramer, M. S. and S. H. Shapiro (1984, November). Scientific challenges in the application of randomized trials. *JAMA: the Journal of the American Medical Association 252*(19), 2739–2745.

LaLonde, R. J. (1986, September). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review 76*(4), 604–620.

Lucas, R. E. and T. J. Sargent (1981). *Rational Expectations and Econometric Practice*. Minneapolis: University of Minnesota Press.

Marschak, J. (1953). Economic measurements for policy and prediction. In W. C. Hood and T. C. Koopmans (Eds.), *Studies in Econometric Method*, pp. 1–26. New Haven, CT: Yale University Press.

Orcutt, G. H. and A. G. Orcutt (1968, September). Incentive and disincentive experimentation for income maintenance policy purposes. *American Economic Review 58*(4), 754–772.

Palca, J. (1989, October 6). AIDS drug trials enter new age. *Science, New Series 246*(4926), 19–21.

Ramey, C. T., A. M. Collier, J. J. Sparling, F. A. Loda, F. A. Campbell, D. A. Ingram, and N. W. Finkelstein (1976). The Carolina Abecedarian Project: A longitudinal and multidisciplinary approach to the prevention of developmental retardation. In T. Tjossem (Ed.), *Intervention Strategies for High-Risk Infants and Young Children*, pp. 629–655. Baltimore, MD: University Park Press.

Rawlins, M. (2008). *De testimonio*: on the evidence for decisions about the use of therapeutic interventions. *The Lancet 372*(9656), 2152 –2161.

Royal Academy of Sciences (2019, October). The prize in economic sciences 2019. Press Release.

Rubin, D. B. (1978, January). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics 6*(1), 34–58.

Sanson-Fisher, R. W., B. Bonevski, L. W. Green, and C. D'Este (2007). Limitations of the

randomized controlled trial in evaluating population-based health interventions. *American Journal of Preventive Medicine 33*(2), 155 –161.

Savage, L. J. (1962). *The Foundations of Statistical Inference: A Discussion Opened by L.J. Savage at the Meeting of the Joint Statistics Seminar, Birkbeck and Imperial Colleges, in the University of London*. New York: Barnes and Noble.

Sehon, S. R. and D. E. Stanley (2003). A philosophical analysis of the evidence-based medicine debate. *BMC Health Services Research 3*(1), 14.

Shahar, E. (1997). A Popperian perspective of the term 'evidence-based medicine'. *Journal of Evaluation in Clinical Practice 3*(2), 109–116.

Silvey, S. D. (1980). *Optimal Design: An Introduction to the Theory for Parameter Estimation*. New York: Chapman and Hall.

Tinbergen, J. (1956). *Economic Policy: Principles and Design*. Amsterdam: North Holland Publishing Company.

Worrall, J. (2007). Why there's no cause to randomize. *The British Journal for the Philosophy of Science 58*(3), 451–488.

Wu, D. (1973, July). Alternative tests of independence between stochastic regressors and disturbances. *Econometrica 41*(4), 733–750.