Journal of Economic Perspectives—Volume 9, Number 2—Spring 1995—Pages 85–110

Assessing the Case for Social Experiments

James J. Heckman and Jeffrey A. Smith

Recent academic debates pit two alternative approaches to policy evaluation against one another. The first is the "experimental" approach, based on the random assignment of accepted program applicants to a recipient, or treatment, group and a non-recipient, or control, group. The second is the "nonexperimental," or "econometric," approach that uses a variety of microdata sources, statistical methods, and behavioral models to compare the outcomes of participants in social programs with those of nonparticipants. The central question addressed in this paper is whether or not randomized social experiments aid in securing answers to basic questions about the evaluation of social programs.

There are many distinct and complementary approaches to the study of the impact of public policy, including full general equilibrium analysis of policy impacts (Tinbergen, 1956; Auerbach and Kotlikoff, 1987; Shoven and Whalley, 1992; Kydland and Prescott, 1991) and less ambitious partial equilibrium microeconomic structural research programs, such as those designed to estimate the impact of taxes on labor supply. Both approaches offer answers to many interesting counterfactual policy questions, but their credibility rests critically on the quality of the empirical input used to generate their answers

■ *James J. Heckman is Henry Schultz Distinguished Service Professor of Economics and Director of the Center for Social Program Evaluation at the Harris School of Public Policy Studies, University of Chicago, Chicago, Illinois. Jeffrey A. Smith is Assistant Professor of Economics at the University of Western Ontario, London, Ontario, Canada, and an Affiliated Faculty Member of the Center for Social Program Evaluation, University of Chicago, Chicago, Illinois.*

and on the widespread acceptance of crucial identifying assumptions. These are often dubious and controversial. As a result, there remains considerable interest in the answers to much more modest, but still very hard, first questions: do social programs have any impacts on participants and, if so, what are they? These are surely the first questions to answer before more elaborate structural models are fit—or imposed—on the data. They are the questions considered in this paper.

In discussing social experiments, we initially confine our attention to the recent black-box version of the experimental method that aims solely at obtaining reliable estimates of the mean impacts of particular programs or treatments. The earlier view that inspired the negative income tax experiments saw experimentation as a tool for obtaining reliable estimates of the parameters of invariant structural models of behavior (for an example, see Orcutt and Orcutt, 1968). We believe the older approach is more likely to produce long-run knowledge. However, in the public policy community there is a widespread perception that structural models are unable to explain behavior. The new emphasis is on determining whether specific programs "work," in the sense of having a positive mean impact, rather than on learning about structural parameters (such as labor supply elasticities) that might be used to evaluate a variety of programs, including some that have never been put in operation.

We illustrate our general arguments regarding social experimentation with empirical evidence from the recently completed experimental evaluation of the training programs provided under Title II-A of the Job Training Partnership Act (JTPA). The JTPA program provides classroom training in occupational skills, basic education, subsidized on-the-job training at private firms, and job search assistance to the disadvantaged. The experimental evaluation was funded by the U.S. Department of Labor and conducted by two leading firms in the field of experimental evaluation, the Manpower Demonstration Research Corporation (MDRC) and Abt Associates. Our discussions of the JTPA experiment draw on their reports (Doolittle and Traeger, 1990; Bloom et al., 1993), as well as on our own work cited below.

In the next section, we present the strongest case for experiments—that they provide a simple solution to the problem of selection bias that nonexperimental analyses must overcome by using econometric methods. We then criticize four other arguments commonly advanced in favor of experimentation. The remainder of the paper reviews the theoretical and empirical case against social experiments. We show that experimental data provide no answers to many of the questions of interest to program evaluators, and present empirical evidence on the failure of key assumptions required to justify experimental estimates. We conclude with a summary and a call for rethinking the current emphasis on black-box experimental analyses, whether in place of nonexperimental analyses or of experiments devoted to obtaining estimates of structural economic models.

How Social Experiments Solve the Evaluation Problem

The strongest argument in favor of experiments is that under certain conditions they solve the fundamental evaluation problem that arises from the impossibility of observing what would happen to a given person in both the state where he or she receives a treatment (or participates in a program) and the state where he or she does not. If a person could be observed in both states, the impact of the treatment on that person could be calculated by comparing his or her outcomes in the two states, and the evaluation problem would be solved. More formally, suppose that a person can be in either a treated state, denoted state "1," or an untreated state, denoted state "0," and that there are outcomes, denoted Y_1 and Y_0, associated with each state. These outcomes might consist of earnings or employment in the two states. The gain (or loss) from treatment, call it Δ, equals the difference in outcomes between the two states, or $Y_1 - Y_0$.

Because we cannot determine the impact of treatment on particular individuals, evaluators focus their attention on the distribution of impacts across persons, or $F(\Delta)$, or on certain features of this distribution. In particular, the expected gain to a randomly selected person in the population, denoted $E(\Delta) = E(Y_1 - Y_0)$, where $E(\cdot)$ refers to the expected value or population average of the quantity inside the parentheses, often constitutes the parameter of interest. For programs that include the entire population, such as social security reform, this parameter gives the information necessary to perform a benefit-cost analysis when combined with information on average costs. For programs that serve only volunteers, such as most job training programs, or programs targeted to certain groups, it makes sense to focus instead on what happens to those who actually participate. Letting $d = 1$ indicate participation and $d = 0$ indicate nonparticipation, we can write the distribution of gains (or losses) for participants as $F(\Delta|d = 1)$ and the expected impact for participants as $E(\Delta|d = 1) = E(Y_1 - Y_0|d = 1)$.[1]

Existing evaluations focus almost exclusively on estimating mean impacts, even though many other aspects of the distribution of gains (or losses) from a program are also of interest. Examples include the median impact of a program and the fraction of persons with a positive impact from participation in a program. We focus on means in this section to make the case for experimentation as strong as possible. In a later section we discuss the information provided by experimental data about other parameters of interest to evaluators.

The difficulty with estimating the mean impact of a program, either for an entire population or for the population of participants in a voluntary program, arises in constructing the desired counterfactual. Consider the case of a

[1]For simplicity, we ignore the dependence of these measures on explanatory variables ("X" variables). The entire analysis can be regarded as conditional upon them.

voluntary program. In order to estimate the mean impact of participation on participants, we need an estimate of the mean outcome that would have been obtained had the participants not participated. Formally, we need to estimate $E(Y_0|d = 1)$. This is more difficult than it seems, because in general we cannot use the mean outcome among nonparticipants, given by $E(Y_0|d = 0)$, as a proxy for what would have happened to participants had they not participated. To see why, note that subtracting the mean outcome among nonparticipants from the mean outcome of participants, $E(Y_1|d = 1) - E(Y_0|d = 0)$, yields

$$\{E(Y_1|d = 1) - E(Y_0|d = 1)\} + \{E(Y_0|d = 1) - E(Y_0|d = 0)\}.$$

While the first term in curly brackets represents the parameter of interest, the second term represents the selection bias caused by the fact that nonparticipants differ from participants in the nonparticipation state. This selection bias term generally does not equal zero. For example, if persons elect to participate in a program precisely because of the poor alternatives available to them outside the program, nonparticipants will have outcomes higher than those that participants would have had if they had not participated, implying a negative selection bias term.

Randomized social experiments solve the problem of selection bias for means by generating an experimental control group composed of persons who would have participated but who were randomly denied access to the program or treatment. Under the assumptions (discussed at greater length below) that randomization does not alter the pool of participants or their behavior and that close substitutes for the experimental treatment are not readily available, the mean outcome of the experimental control group estimates the desired counterfactual, $E(Y_0|d = 1)$. Let $d^* = 1$ for persons who would participate in a program in the presence of random assignment and $d^* = 0$ for everyone else. Randomization is applied to the population for whom $d^* = 1$. Let $r = 1$ denote randomization into the treatment group, and let $r = 0$ denote randomization into the control group, which is denied access to the treatment. Under the assumptions just described, it follows that the outcomes of the experimental treatment group measure the normal outcomes of program participants, so that $E(Y_1|d = 1) = E(Y_1|r = 1$ and $d^* = 1)$, and that the outcomes of the experimental control group measure what the participants' outcomes would have been had they not participated, so that $E(Y_0|d = 1) = E(Y_0|r = 0$ and $d^* = 1)$. We can then write the mean impact of treatment on the treated for a voluntary program as the difference between the two means:

$$E(Y_1 - Y_0|d = 1) = E(Y_1|r = 1 \text{ and } d^* = 1) - E(Y_0|r = 0 \text{ and } d^* = 1).$$

The mean outcomes of the experimental treatment and control groups provide estimates of the two terms on the right-hand side. Thus, as shown in Heckman (1993a, b) and Heckman and Roselius (1994), randomization acts as an instru-

mental variable by creating variation in the receipt of treatment among participants. This equation also reveals the importance to the case in favor of social experiments of the assumption that mean impacts are the primary parameters of interest. For parameters such as the median impact, which depend on the joint distribution of outcomes in the treated and untreated states, this simple relationship does not apply.[2]

One simplified special case dominates thinking in the evaluation literature. Called the "common-effect" case, it refers to the situation where everyone has the same gain (or loss) from a program, so that $\Delta = (Y_1 - Y_0)$ is the same for everyone. It particularly favors social experiments for two reasons. First, even if randomization changes the pool of persons being treated the parameter being estimated remains the same. Second, in this special case, the link between outcomes in the two states is known for each individual regardless of their observed state. Under this assumption, experimental data provide the full joint distribution of outcomes and so allow estimation of other parameters of interest, such as the median impact and the effect of universal participation. Formally, in this special case $E(\Delta|d = 1) = E(\Delta|d^* = 1) = E(\Delta) = \Delta$. The more general (and more realistic) case where the impact of treatment varies across persons corresponds to the econometric "random-coefficient" model (Heckman and Robb, 1985; Heckman, 1992).

Finally, note that random assignment does not remove selection bias, but instead balances the bias between the participant and nonparticipant samples. To see this, consider the simple common coefficient model: $Y = \alpha + \beta d + U$, where Y is some outcome of interest, α is the mean outcome when no one participates, β is the common effect of participation, and U represents a random shock observed by the individual but not by the analyst. In this model, $\Delta = \beta$, a constant. The selection problem arises when participation, indicated by d, depends on the unobserved random shock U, so that $E(U|d) \neq 0$. Mean earnings in the experimental treatment and control groups are

$$E(Y|r = 1, d = 1) = \alpha + \beta + E(U|d = 1)$$

and

$$E(Y|r = 0, d = 1) = \alpha + E(U|d = 1),$$

respectively. Subtracting the two means yields β, the parameter of interest. Nothing about randomization guarantees that $E(U|d^* = 1) = 0$ or $E(U|d = 1) = 0$. Rather, randomization balances the bias in the two samples, so that it cancels out when calculating the mean impact estimate. (This argument is developed more generally in Heckman and Roselius, 1994; and Heckman, Ichimura, Smith and Todd, 1995.)

[2]The median of the treatment group minus the median of the control group does not, in general, estimate the median gain. Without additional assumptions, the median impact cannot be estimated from experimental data.

Arguments for Social Experiments

In the next main section, we show that while selection bias can affect nonexperimental analyses, experiments can induce biases of their own. In this section, we review and critique four weaker arguments commonly advanced in support of experimental methods.

The Selection Problem is Universal, and Nonexperimental Methods Cannot Solve It

The empirical case demonstrating the importance of selection bias and the fragility of the nonexperimental methods used to evaluate social programs relies heavily on LaLonde (1986). LaLonde uses an experimental evaluation of the National Supported Work Demonstration (NSW) as a benchmark against which to compare nonexperimental estimates. He uses the NSW experimental treatment group in conjunction with comparison groups drawn from the Current Population Survey (CPS) and the Panel Study of Income Dynamics (PSID) to estimate the impact of training. He employs several commonly used nonexperimental estimators and obtains a wide variety of impact estimates, most of which differ substantially from the corresponding experimental estimates. A limited set of model selection tests fails to eliminate the models that produce this variability.[3]

This study has had a strong influence in promoting the use of experiments to evaluate social programs in general, and employment and training programs in particular (for example, Hansen, 1994, p. 101). Despite its influence, this study has important limitations that serve to limit the generality of its methodological conclusions.

Selection bias arises because of missing data on the common factors affecting participation and outcomes. The most convincing way to solve the selection problem is to collect better data. This option has never been discussed in the recent debates over the merits of experimental and econometric approaches and has only recently been exercised. Heckman, Ichimura, Smith and Todd (1995) and Heckman and Roselius (1994) demonstrate that sufficiently rich data collected on persons who are both eligible for JTPA and are located in the same labor markets can be used to create a nonexperimental comparison group that is virtually identical to the control group from the recent JTPA experiment. The data used by LaLonde (1986) either lack sufficient information to determine eligibility for the NSW program or the use of eligibility was not considered as a screening criterion in forming comparison groups. Furthermore, sample sizes are too small, and insufficient geographical information is available in LaLonde's data to place comparison group members in the same

[3]Fraker and Maynard (1987) use a similar strategy to evaluate alternative comparison group designs. Their study has the same limitations as those in LaLonde (1986).

labor markets as program participants. In short, the problem of selection bias documented by LaLonde (1986) arises at least in part from the crudity of his data.

A second limitation of the data available on the NSW participants and on comparison group members from the CPS and PSID is that many of the nonexperimental estimators developed in the literature cannot be applied to them. For example, the data on NSW participants contains only a single year of preprogram earnings information, effectively ruling out the use of many estimators based on the longitudinal structure of earnings discussed in Heckman and Robb (1985). In addition, the relative paucity of conditioning variables or regressors in the CPS and the PSID rules out effective strategies for controlling for unobservables by including a rich set of observables. LaLonde's study also fails to address the choice-based nature of his sample, which affects the properties of many of the estimators he examines (Heckman and Robb, 1985). As a result, inappropriate application of certain econometric methods causes part of the variability he finds.

A third factor limiting the generalizability of the findings from this study is its failure to utilize a variety of model-selection strategies based on standard specification tests. Heckman and Hotz (1989) reanalyze the NSW data used by LaLonde and find that a simple set of specification tests successfully eliminates all but the nonexperimental models that reproduce the inference obtained by experimental methods. They conclude that specification tests remain a promising tool for nonexperimental analysts. Heckman (1993b) and Heckman and Roselius (1994) discuss the limitations of these tests.

A fourth limiting factor is that LaLonde's study treats the choice of a comparison group and the choice of an estimator as statistical, rather than economic, problems. As a result, he ignores the potential of cumulative social science knowledge to guide these choices. Recent years have witnessed the accumulation of substantial empirical knowledge in areas like the dynamics of individual earnings and the process of selection into social programs. This knowledge is sufficient to rule out in advance some commonly used estimators of program impact, such as the fixed effect or "difference-in-differences" estimator that is almost always rejected in applications of specification tests to nonexperimental data (Heckman, 1993b). By ignoring the evidence available from cumulative social science knowledge, LaLonde (1986) ends up testing only a weak and incomplete version of nonexperimental methodology.

A final factor limiting the generalizability of the LaLonde (1986) study is that since that paper was written, there have been important developments in nonexperimental evaluation methods. Real progress has been made in relaxing the strong distributional and functional form assumptions maintained in the earlier literature on controlling for sample selection bias. These assumptions were the target of frequent criticism. Examples from the large recent literature on semiparametric and nonparametric estimation include Andrews (1991), Cosslett (1991), Ichimura and Lee (1991), Newey (1988), Powell (1989) and the

overviews in Heckman (1990a, b, 1993b). The development of these methods challenges the strong methodological conclusions reached by Lalonde and others regarding the ineffectiveness of nonexperimental methods.

Experiments Are Based on More Plausible Assumptions

Both experimental and nonexperimental evaluation methods face the same fundamental problem that no person is observed simultaneously in both the treated and untreated states. Many methods have been proposed to construct the counterfactual outcome corresponding to what persons who did participate in a program would have done had they not participated. In an experiment, the counterfactual is represented by the outcomes of a control group generated through the random denial of services to persons who would ordinarily be participants. In a nonexperimental setting, the counterfactual is obtained econometrically, using models of the program participation and outcome processes. Both settings require assumptions, but proponents of experiments argue that experiments require fewer or more plausible assumptions than do nonexperimental evaluations. In this section, we compare the assumptions required for an experiment to those required for a nonexperimental evaluation.

As previously noted, for the outcomes of an experimental control group to correspond to the outcomes that participants would have experienced had they not participated in the program, two assumptions must hold. The first assumption requires that randomization not alter the process of selection into the program, so that those who participate during an experiment do not differ from those who would have participated in the absence of an experiment. Put simply, there must be no "randomization bias." Under the alternative assumption that the impact of the program is the same for everyone (the conventional common-effect model), the assumption of no randomization bias becomes unnecessary, because the mean impact of treatment on participants is then the same for persons participating in the presence and in the absence of an experiment.[4]

The second assumption is that members of the experimental control group cannot obtain close substitutes for the treatment elsewhere. That is, there is no "substitution bias." In the presence of substitution bias, the experimental control group no longer corresponds to the desired counterfactual of persons who wanted to receive treatment but did not, and as a result the mean difference in outcomes between the treatment and control groups no longer provides an estimate of the mean impact of treatment on the treated. We present empirical evidence on the validity of these assumptions below.

The basic nonexperimental assumption is that a model of the outcome process can be determined, along with the relationship between the outcome

[4] Note that randomization bias, which results from different patterns of participation in the presence of an experiment, differs from a "Hawthorne effect." The latter arises from the act of observation itself and can occur in either experimental or nonexperimental evaluations.

process and the process of selection into the program being evaluated.[5] For example, in the case of an employment and training program, a model of the earnings behavior of the population served by the program must be determined from the available data and from economic theory, and the effect of earnings on selection into the program must be determined. Unlike experimental evaluations, nonexperimental evaluations can build on cumulative knowledge about earnings and selection processes from prior studies, as well as information about features of the outcome and selection processes that can be gained from the data at hand.

At this level of generality, both sets of assumptions are simple to understand and plausible.

Experimental Results are Easier to Explain to Policymakers

It has been argued that experimental evidence on program effectiveness is easier for politicians and policymakers to understand. This argument mistakes apparent for real simplicity. In the presence of randomization bias or substitution bias, the meaning of an experimental impact estimate would be just as difficult to interpret honestly in front of a congressional committee as any nonexperimental study. The hard fact is that some evaluation problems have intrinsic levels of difficulty that render them incapable of expression in sound bites. Delegated expertise must therefore play a role in the formation of public policy in these areas, just as it already does in many other fields. It would be foolish to argue for readily understood but incompetent studies, whether they are experimental or not.

Moreover, if the preferences and mental capacities of politicians are to guide the selection of an evaluation methodology, then analysts should probably rely on easily understood and still widely used before-after comparisons of the outcomes of program participants. Such comparisons are simpler to explain than experiments, because they require no discussions of selection bias and the rationale for a control group. Furthermore, before-after comparisons are cheaper than experiments. They also have the advantage, or disadvantage, depending on one's political perspective, that they are more likely to yield positive impact estimates (at least in the case of employment and training programs) due to the well-known preprogram dip in mean earnings for participants in these programs (Ashenfelter, 1978; Heckman and Smith, 1994a).

Most advocates of social experiments would reject replacing them with before-after comparisons because the implicit assumptions underlying such comparisons, such as the absence of economy-wide factors affecting participant outcomes, often fail to hold in practice. The same concern about the validity of key assumptions should also apply to experiments.

[5]In more particular terms, the list of assumptions justifying nonexperimental methods is immense. Heckman and Robb (1985) categorize and justify a wide variety of these assumptions.

Furthermore, policymakers often do not care solely about whether or not a particular program "works" in the sense of having benefits that exceed its costs. When programs fail, it is important to understand why they do not work. Without this information, which is not available from typical black-box experimental analyses, the only alternative open to politicians is to eliminate one program completely and start fresh with another. With the additional information available through nonexperimental methods, or through experiments designed to uncover the parameters of invariant structural models, it can be determined which services offered by a program work and for whom, thus allowing politicians to retarget and redesign existing programs, whose effectiveness will increase along with the store of social science knowledge about them.[6]

Experiments Produce a Consensus

A final argument offered in favor of experiments is that they produce "one number" rather than the bewildering array of nonexperimental estimates often found in the literature on program evaluation. In assessing this argument, it is important to distinguish the consensus produced by monopoly from the consensus that emerges from scholarship. Many organizations producing experimental analyses have been unwilling to share their data with the academic research community. The appearance of a consensus view is a consequence of only one interpretation of the data being given.[7]

The analyses of the effect of transfers on marital dissolution using the SIME-DIME (Seattle-Denver Income Maintenance Experiment) experimental data provide a good example of this point. When the SIME-DIME data became publicly available, strong disagreements emerged over the interpretation of the experimental evidence (Hannan and Tuma, 1990; Cain and Wissoker, 1990). Earlier analyses of the effects of negative income taxes on labor supply provoked similar controversy.

Criticisms of Social Experiments

The limitations of nonexperimental methods are the topic of the rich and active field of econometrics. The limitations of experimental methods have received less critical scrutiny in the research community. One should not confuse unexamined presumptions about experiments with actual knowledge about the importance of the biases induced by experimentation. In this portion of the paper, we turn our attention to five major criticisms that have been made

[6]It has also been argued that in comparison to nonexperimental methods, experiments permit measurement of the effects of new kinds of treatment that have not previously been observed (Burtless, 1993). This argument confuses demonstrations—temporary implementations of new programs for research purposes—with experiments. Demonstrations may or may not be effectively evaluated with nonexperimental methods.

[7]MDRC, one of the major producers of experimental evaluations, has recently announced that it will begin to make its experimental data publicly available.

of experimental methods for evaluating social policies. For brevity, we omit some criticisms that have been treated in detail elsewhere, such as ethical objections to random assignment (for example, Burtless and Orr, 1986); the long delays often associated with experimental evaluations;[8] attrition from experimental samples (Hausman and Wise, 1985); and the inability of small-scale experiments to predict general equilibrium effects or to produce results that can be extrapolated to other populations (Zellner and Rossi, 1986). Elsewhere, we discuss how economic theory can be used to supplement experimental data to allow generalizations to other populations (Heckman and Smith, 1993, 1995).

Experiments Provide Little Evidence on Many Questions of Interest

There are many questions of interest to program evaluators. Earlier, we noted that the case for social experiments rests strongly on the idea that the primary objects of interest consist of estimates of the mean impact of a program or treatment on either the population as a whole (for mandatory programs) or on participants (for voluntary programs). In a later section, we discuss the ability of experimental data to provide information about other aspects of the distribution of program impacts, such as the median gain or the fraction of participants who benefit from a program. Here, we note that many important evaluation questions do not involve the distribution of impacts at all.

Policymakers care about the answers to at least four other questions not addressed in black-box experimental studies that focus only on obtaining impact estimates. These questions are as follows: What are the effects of factors such as subsidies, advertising, local labor markets, family income, race and sex on program application decisions (Heckman, 1992; Moffitt, 1992)? What are the effects of bureaucratic performance standards, local labor markets and individual characteristics on administrative decisions to accept applicants and place them in specific programs? What are the effects of family background, subsidies and local market conditions on decisions to drop out from a program and on the length of time taken to complete a program? What are the costs of various alternative treatments?

Some of these questions might in principle be evaluated using random assignment designs, but practical difficulties would make it impossible in most cases. For example, while subsidies for program entry or completion could in principle be randomly assigned, family background variables and local labor market conditions cannot be. Since experiments can answer only a subset of the questions of interest to evaluators, it remains important to build up the stock of basic social science knowledge required to successfully utilize nonexperimental methods, both by themselves and as a tool for more extensive analyses of experimental data.

[8]For example, the recent experimental JTPA evaluation took eight years to conduct and presents estimates for a program that has been fundamentally altered since the evaluation's data were collected.

The Intrinsic Variability in Evidence from Randomized Experiments

As discussed above, experiments provide estimates of the mean difference in outcomes between persons receiving and not receiving some treatment. In this section, we provide evidence on the intrinsic variability present in experimental data and on the inability of experimental data to provide useful information about the overall distribution of program impacts (unless aided by further assumptions or prior information). The analysis in this section draws on Clements, Heckman and Smith (1993) and Heckman and Smith (1995).

The data obtained from an experiment consists of two marginal distributions of outcomes, $F(Y_1|d = 1)$ and $F(Y_0|d = 1)$, one for those in the treatment state and one for those in the control state. These distributions are sufficient to identify a number of parameters of interest, including the mean impact of treatment, and with some additional information about the utility function, expected utility in the treated and untreated states. In addition, the budgetary impact of a program can be estimated by combining information on earnings outcomes with information on tax schedules and program costs.

However, because we do not observe anyone in both the treated and the untreated states, experimental data do not provide the joint distribution of outcomes in the two states. That is, they do not indicate the (probabilistic) relationship between outcomes in the two states. In the special case of the common effect model, social experiments do identify the full joint distribution. In this special case, $Y_1 - Y_0 = \Delta$, a constant, for everyone, where Δ can be identified using experimental data. Given Δ, knowledge of either Y_1 or Y_0 determines the other. Graphically, the distribution of Y_1 equals the distribution of Y_0 shifted over by Δ. More formally, if F_1 is the cumulative distribution function (CDF) of Y_1 and F_0 is the CDF of Y_0, then $F_1(Y_0 + \Delta) = F_0(Y_0)$.

Identification of many other parameters of interest requires knowledge of the full joint distribution. For example, policymakers concerned with equity (and with reelection) would like to know the fraction of persons made better off by a program, along with various quantiles of the impact distribution. In modelling individual participation choices, knowledge of the joint distribution is required when agents are assumed to know their outcome in one of the two states but not the other, for in this case their decisions depend on the distribution of outcomes in one state conditional on the known outcome in the other. (See Heckman and Smith, 1995.)

For simplicity, consider the case where the outcome variable is discrete, such as employment. Those who are randomized into a program may be employed or not employed after completing it. Those who are randomized out of a program may also be employed or not employed in the evaluation period. The latent distribution underlying this situation is a bivariate binomial. Let (E, E) denote the event "employed with treatment and employed without treatment," and let (E, N) denote the event "employed with treatment, not employed without treatment." Similarly, (N, E) and (N, N) refer respectively to cases where a person would not be employed if treated but would be employed if not treated and where a person would not be employed in either case. The

Figure 1
A Contingency Table

		Untreated: E	Untreated: N	
Treated	E	P_{EE}	P_{EN}	$P_{E\bullet}$
	N	P_{NE}	P_{NN}	$P_{N\bullet}$
		$P_{\bullet E}$	$P_{\bullet N}$	

probabilities associated with these events are denoted by P_{EE}, P_{EN}, P_{NE} and P_{NN}, respectively.

This model of outcomes appears in the form of a contingency table in Figure 1. The columns refer to employment and nonemployment in the untreated state. The rows refer to employment and nonemployment in the treated state. If we were able to observe the same individuals in both the treated and untreated states, we could fill in the table and estimate the full distribution of program outcomes for everyone. Instead, from randomized trials we can estimate row and column totals. That is, we can estimate $P_{E\bullet}$ (the sum across the top row) using the employment proportion among those treated, which consists of those who would not have had jobs without the treatment, and those who would have had jobs anyway. Similarly we can estimate $P_{\bullet E}$ (the sum of the first column) using the employment proportion among the untreated, which consists of those who would have been employed with or without treatment, and those who would *not* be employed if treated, but would be employed if not treated.

The impact of treatment is defined as $T = P_{EN} - P_{NE}$, the proportion of people who would switch from nonemployed to employed as a result of treatment minus the proportion of persons who would switch from being employed to nonemployed as a result of the treatment. This is a *net* measure of the impact of treatment. From the contingency table, it is evident that $T = P_{E\bullet} - P_{\bullet E}$. Thus, T can be estimated without bias by subtracting the proportion employed in the control group ($P_{\bullet E}$) from the proportion employed in the treatment group ($P_{E\bullet}$). Experimental data provide exactly the information required to estimate T.

If we wish to decompose T into its two components, however, the experimental data do not give an exact answer except in special cases. Such a decomposition is of interest if we seek to learn the extent to which the program actually harmed participants, as indicated by the magnitude of P_{NE}. In terms of the contingency table, we know the row and column marginals, but not the individual elements in the table.

Frechet (1951) and Hoeffding (1940) demonstrate how to bound joint distributions from knowledge of the marginal distributions. The intuition

Table 1
Employment Percentages and Bounds on the Probabilities P_{EN} and P_{NE}

	Adult Males	*Adult Females*	*Male Youth*	*Female Youth*
% Employed: Treatment	0.72	0.64	0.74	0.57
% Employed: Control	0.71	0.61	0.77	0.58
Bounds on P_{EN}	[.01, .29]	[.03, .39]	[.00, .23]	[.00, .42]
Bounds on P_{NE}	[.00, .28]	[.00, .36]	[.03, .26]	[.01, .43]

Notes: Employment Percentages are based on percentage employed in months 16, 17 and 18 after random assignment. P_{ij} is the probability of having employment status i as a treatment and employment status j as a control, where i and j take on the values of N and E. The Frechet-Hoeffding bounds are then given by

$$P_{ij} \leq \text{FUB}(P_{ij}) = \min\{P_{Nj} + P_{Ej}, P_{iN} + P_{iE}\} \text{ and}$$

$$P_{ij} \geq \text{FLB}(P_{ij}) = \max\{[P_{Nj} + P_{Ej}] + [P_{iN} + P_{iE}] - 1, 0\}.$$

behind these bounds is simple. The upper bound results from the fact that the probability of a joint event can never exceed the probability of the events that compose it. Thus, for example, the probability of being employed in both states, P_{EE}, cannot exceed the smaller of the two probabilities in the individual states, $P_{E\bullet}$ and $P_{\bullet E}$. The lower bound results from the condition that the sum of the four individual cell probabilities must equal one. To see how this provides a bound, suppose that $P_{E\bullet}$ and $P_{\bullet E}$ both equal 0.6. This implies that the element of the contingency table that they have in common, P_{EE}, must equal at least 0.2 or the sum of the probabilities in the table will exceed 1.0. If P_{EE} takes on a smaller value, say 0.1, then P_{NE} and P_{EN} would both equal 0.5, and the sum of the individual probabilities would equal 0.5 + 0.5 + 0.1 = 1.1, which exceeds 1.0.

How wide are the ranges implied by these bounds? To address this question, we calculate them using data from the JTPA experiment. Table 1 presents the Frechet-Hoeffding bounds for P_{NE} and P_{EN}.[9] They are very wide. Even without taking into account sampling error, the experimental evidence for adult males is consistent with P_{EN} (the fraction employed with treatment but not employed without it) ranging from 0.01 to .29. The range for P_{NE} (the fraction not employed if treated, but employed without treatment) is equally large. As many as 28 percent and as few as none may have their employment prospects diminished by participating in the program. The two probabilities

[9]Employment is defined as positive, self-reported earnings in the 16th, 17th or 18th months after random assignment.

are not functionally independent; because T is known and $T = P_{EN} - P_{NE}$, it follows that high values of P_{EN} are associated with high values of P_{NE}.

From this evidence, we cannot distinguish between two stories: that the JTPA program benefits many people in terms of facilitating their employment but also harms many people in that they are less likely to work than if they had not participated, or that the program benefits and harms only a small percentage of those it serves.

Similarly wide bounds emerge from an examination of the earnings data produced by the JTPA experiment. For adult men, the correlation between earnings (over the 18 months after random assignment) in the treatment and control states is bounded between -0.79 and 1.00, and the standard deviation of the impact of the program (defined as the difference in earnings between the treatment and control states) is bounded between \$821 and \$21,857. For all four demographic groups, this important parameter is bounded away from zero, indicating that the data reject a model of equal program impacts across persons. Using related techniques, we find that for adult men the 25th percentile of the impact distribution has a range of at least $-\$15{,}500$ to \$200, while the 75th percentile has a range of at least \$900 to \$16,700. We find similarly wide bounds for adult women and for male and female youth. In each case, conditioning on the available regressors does *not* substantially reduce the range of variability in these estimates. There is considerable uncertainty about policy-relevant parameters estimated from *ideal* experimental data. Clements, Heckman and Smith (1993) present methods for incorporating prior information into the analysis of experimental data to reduce the uncertainty inherent in them.

These calculations demonstrate the variability intrinsic in data from a social experiment. Only if the evaluation problem is defined exclusively in terms of means can it be said that experiments provide a precise answer. Experiments fail to provide clear and convincing evidence of the effect of treatment on many interesting features of the outcome distribution.

Randomization Bias

Randomization bias occurs when random assignment causes the type of persons participating in a program to differ from the type that would participate in the program as it normally operates. Randomization bias also results from changes in participant behavior due to the threat of service denial, like reductions in complementary training activities undertaken prior to application to the program. Surprisingly, little is known about the empirical importance of randomization bias. Except for the JTPA evaluation, randomized social experiments have only been implemented for demonstration projects designed to evaluate new programs. The possibility of disruption by randomization cannot be confirmed or denied on data from these experiments because there are no nonexperimental versions of these programs.

A recent report by MDRC (Doolittle and Traeger, 1990), based on their experience in implementing the experimental evaluation of JTPA, provides

suggestive evidence on the practical importance of randomization bias. (See also the discussion in Hotz, 1992.) Job training under JTPA is organized through geographically decentralized training sites, whose participation in the experiment was not compulsory. In attempting to enroll geographically dispersed sites, MDRC experienced a training center refusal rate in excess of 90 percent. The reasons for site refusal to participate in the experiment are given in Table 2. (The reasons stated there are not mutually exclusive.) Ethical and public relations concerns lead the list of objections to randomization. Sites expressed major fears (items 2 and 3) about the effects of randomization on the quality of the applicant pool. A lowering of the quality of the applicant pool could impede center performance and thereby reduce the incentive payments they receive based on trainee performance under the JTPA performance standards system.

To form an experimental control group, centers had to expand the set of persons deemed acceptable for the program. This is precisely the behavior that creates randomization bias. To recruit the additional applicants needed to fill the control group while holding constant the number of persons trained, some sites made substantial changes in their recruiting and intake procedures and thus changed the composition of their trainee pool. The MDRC analysts conclude (Doolittle and Traeger, 1990, p. 121):

> 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.

Randomization also creates controversy in clinical trials analysis in medicine, which is sometimes held up as a paragon for empirical social science (Ashenfelter and Card, 1985). Writing in the *Journal of the American Medical Association*, Kramer and Shapiro (1984) note that subjects in drug trials were less likely to participate in randomized studies than in nonexperimental studies. They discuss one study of drugs administered to children afflicted with a disease. The study had two components. The nonexperimental part 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. These authors cite evidence suggesting selective failure to participate in randomized trials. In a study of the 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.

Table 2
Percent of Training Centers Citing 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

Source: Based on the responses of 228 JTPA 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. Percentages may add to more than 100.0 because training centers could raise more than one concern.

The evidence suggests that randomization bias is not just a theoretical issue. Instead, it is an empirically important problem in both social experiments and clinical trials in medicine that is usually ignored by advocates of social experimentation.

Institutional Limitations on Social Experiments

The institutional structure of social programs places limits on the use of randomized social experiments. In this section, we discuss three important limitations. First, institutional factors can make it difficult to choose the optimal placement of random assignment within the overall process of program participation. This placement affects attrition from the program within the treatment group and the interpretation of the resulting experimental estimates. Second, institutional factors can make it impossible to produce separate experimental estimates of the impact of individual treatments. Third, voluntary participation of sites in an experiment limits the external validity of the derived estimates. Although our discussion draws primarily on the recent experimental JTPA evaluation, each of these same limitations arises in the evaluation of many other social programs.

The decision to participate in a program can be broken into a series of steps: each participant becomes eligible for the program; becomes aware of the program; realizes his or her own eligibility; applies to the program; is accepted

into the program; is assessed by program staff; is assigned to particular program services (such as classroom training or job search assistance in JTPA); begins receiving treatment; and then ultimately completes the treatment. In theory, randomization can occur at any step in this sequence. The optimal placement depends on the evaluation question being answered. Angrist and Imbens (1991) and Heckman and Smith (1993) discuss the merits of alternative points of randomization.

The experience gained in implementing the JTPA experiment suggests that practical considerations can severely limit the ability of researchers to choose the optimal placement of random assignment. In the JTPA evaluation, the parameter of interest was the mean impact of JTPA training on those receiving it. Given this parameter of interest, random assignment should be placed so as to minimize attrition from the program within the treatment group, because the experimental mean-difference estimate corresponds exactly to the impact of training on the trained only in the case where there is no attrition. Locating random assignment as close as possible to the actual initiation of training reduces the opportunity for dropping out of the program in the period between random assignment and receipt of training.

Abt Associates and MDRC recognized that the initiation of training was the best time for random assignment, but institutional and political factors made it impracticable to choose this point. Instead, random assignment took place just after assignment to a particular service type. While assignment to a particular service and initiation of treatment are consecutive steps in the participation process, in practice they may be separated in time by weeks or even months. As many occupational training classes are offered on an academic schedule, a trainee assigned to a particular training course must often wait until the beginning of the next academic quarter or semester to begin training. Similarly, a person assigned to receive subsidized on-the-job training at a private firm must wait while the training site locates potential training opportunities. During these waiting periods, the JTPA applicant may move out of town, find a job on his or her own, become disinterested in the training offered, or take training from another source. As a result, there is substantial attrition in the experimental treatment group.

Cost considerations led to the unsatisfactory placement of random assignment in the JTPA experiment. The later random assignment occurs in the program participation process, the more resources are spent on applicants who will ultimately end up in the control group. Training centers resist these expenses, as they receive no reward for having to incur them. A further source of costs results from the decentralized nature of the JTPA program, in which assessment and assignment to services typically occur in a central office, while the actual training occurs at the location of the service provider. Placing random assignment earlier in the process avoids the cost of involving the provider in random assignment and compensating the provider for processing the applicant before randomization takes place.

Thus, there are real costs to randomizing later rather than earlier in a program. The failure to locate random assignment optimally in the JTPA evaluation does not represent an accidental oversight in implementation that can be cheaply corrected in subsequent experimental evaluations. In the JTPA experiment, just under 65 percent of the treatment group eventually received JTPA employment and training services, according to the administrative records of the sites themselves. In the presence of this level of attrition, the only way to obtain a credible estimate of the mean impact of the program is to model the attrition process using the very nonexperimental methods eschewed by proponents of randomized social experiments. Heckman, Smith and Taber (1994) offer a discussion of these methods. Hotz and Sanders (1994) describe an alternative approach that generalizes to a multiple treatment environment.

The second institutional problem facing experiments is the difficulty of generating separate experimental estimates of the impact of different service types. For example, the JTPA program offers a number of different employment and training services. Some participants receive a single service type, while others receive specially designed sequences of services. In the JTPA program, obtaining estimates of individual services was one of the primary objectives of the U.S. Department of Labor in commissioning the experiment. However, the structure of the JTPA program makes it impossible to obtain such estimates, at least without multistage randomization or substantial changes in program operation. The problematic features of the JTPA program appear in many other social programs and pose a substantial challenge to the experimental methodology.

The JTPA program is designed to allow local operators wide flexibility in tailoring the treatments offered to the needs and goals of the client. In the absence of an experiment, the process of assignment to particular JTPA services is ongoing. An initial set of service recommendations is made following assessment of the client. The training actually received depends on a number of factors, including the initial recommendations, additional information gained from continued interaction with the client, the availability of classroom training slots, the willingness of private firms to provide on-the-job training, and the amount of funds remaining in the site budget. For example, a client recommended for both on-the-job training and job search assistance might receive the latter while waiting for the former. If the job search assistance is successful, it is the only service received. If it is not, the client goes on to receive on-the-job training.

Due to this flexibility in the system—which is highly valued by program operators—there is no point prior to termination from the program at which the set of services to be received by a given client is known with certainty. Thus, there is no way to obtain experimental estimates of the impact of individual service types unless randomization occurs at more than one point in the participation process. The possibility of using multistage randomization to deal with this problem is not mentioned in MDRC's first implementation report

(Doolittle and Traeger, 1990). It was presumably rejected due to its cost, difficulty of implementation, likely disruptive effect on the program, and the large sample size required to produce meaningful estimates at the final stage. The failure to obtain such estimates is not an accidental feature of the JTPA evaluation; it represents a systematic, institutional limitation on the use of experimental methods.

The great difficulties MDRC experienced in inducing JTPA training sites to participate voluntarily in the JTPA experiment illustrate the third institutional limitation on the practical implementation of experiments. The Department of Labor preferred voluntary site participation over mandated participation for political reasons and because of concerns about whether sites forced to participate would adhere to the experimental protocols without costly monitoring (Doolittle and Traeger, 1990). Since the JTPA experiment was begun, corroborating evidence from Norway has appeared that lends credence to fears regarding the effects of forced site participation (Torp et al., 1993). In that country, training program administrators forced to participate in an experiment successfully undermined random assignment by manipulating their administrative data systems. When asked to randomize 64 persons into 32 slots, administrators used their discretion to declare 31 of the persons ineligible or inappropriate so that the randomization for the 32 slots was actually conducted on a pool of 33 potential participants.

One of us (Heckman) attempted to convince a local JTPA official to use random assignment to allocate persons to limited spaces in a new program. The official vehemently objected, arguing that "there are only a few motivated persons out there, and if I randomize some of them out, I will fail to fulfill my contract to produce trainees at a certain level." Not only are the fears of these officials real, but they have the power to subvert any randomization imposed upon them. In the JTPA experiment, after more than a year of searching and with the help of hundreds of thousands of dollars in side payments, MDRC ended up with a nonrandom sample of 16 sites willing to participate.

As the participating sites are not a random sample, the extent to which the experimental results apply to sites other than those in the experiment remains unknown. In his accompanying paper in this issue, Burtless suggests that this external validity problem could be overcome in future experimental evaluations by forcing most or all JTPA sites to participate, but assigning only a few persons at each site to the control group. This approach would be extraordinarily costly because of the high fixed costs associated with training staff at each site to do random assignment and with monitoring the sites to insure their adherence to experimental protocols. Moreover, it ignores the willingness and ability of unhappy administrators to undo randomization. Finally, it ignores the fact that in decentralized programs such as JTPA, each site is in a meaningful sense a separate program. For such programs, it is useful to perform separate evaluations at a few sites to learn about their individual character, instead of combining a few observations from each of a large number of sites.

Substitution Bias

Substitution bias arises when members of an experimental control group gain access to close substitutes for the experimental treatment, like similar services offered by other providers or the same service offered under different funding arrangements.[10] This situation often arises in clinical trials, when human subjects recognize that they have been denied treatment and attempt to obtain it elsewhere.

In the presence of substitution bias, control group outcomes no longer correspond to the untreated state. To see this, consider the following example. Suppose that half the participants in some program would gain $10,000 from receiving training, while the other half would gain only $100. Participants learn their type after acceptance into the program but prior to receipt of training. Suppose further that the program consists of a subsidy that reduces the effective price of training from $200 to zero. In an experimental setting, everyone in the treatment group would take training, but in the control group only those gaining $10,000 would take training. For the remainder of the control group, the $200 cost exceeds the $100 benefit. The experimental mean-impact estimate (excluding costs) would be $50. This figure does not represent the mean impact of the training on those who received it in either group. What it does represent depends critically on assumptions about how participation decisions for the program and its substitutes respond to the introduction of random assignment.

Substitution bias is particularly important in the evaluation of ongoing, voluntary government programs that provide services also available from (or subsidized by) other government agencies, private nonprofit organizations, and private firms. The JTPA program discussed above is just such a program. Many JTPA training sites contract out with local colleges, vocational schools, and community-based organizations to provide classroom training and other services to JTPA enrollees. Often the JTPA program simply purchases spaces in publicly available training programs. Control group members denied access to a particular program through JTPA can simply purchase a space for themselves or obtain other governmental subsidies such as Pell grants. To make randomization more palatable to the experimental sites, MDRC allowed them to provide a list of alternative service providers to everyone in the experimental control group. While two of the 16 sites did not provide a list, others provided lists many pages in length. In addition, in some sites the JTPA program is co-located with the state employment service, which provides job search assistance to the unemployed. Persons randomized out of JTPA at such sites face a low cost of obtaining these services, which are similar to those provided to many JTPA enrollees.

[10]Crossover bias is a related problem that occurs when control group members gain access to the experimental treatment itself. Evidence from numerous social experiments suggests that careful implementation can prevent this problem. See, for example, Bloom et al. (1993).

The potential empirical importance of substitution bias is demonstrated by the evidence from the JTPA evaluation.[11] As mentioned earlier, administrative records indicate that 65 percent of those in the treatment group actually received treatment in the 18 months following random assignment. According to the self-reports of the treatment group members, however, only 48 percent received treatment during this period. Meanwhile, 32 percent of control group members self-reported receiving training from other sources over the same interval. Among eligible persons not participating in JTPA surveyed at four of the experimental sites, 15 to 24 percent reported receiving training over a similar time period. These figures indicate that a substantial fraction of the control group received training during the period of the experiment. Using the eligible nonparticipants as a benchmark, controls received training at a level well in excess of that normally observed in the low-income population eligible for JTPA.

The consulting firm responsible for analyzing the results from the JTPA experiment, Abt Associates, suggests in Bloom et al. (1993) that in the presence of substitution bias the estimates obtained by comparing treatment group outcomes to control group outcomes can be interpreted as "the *difference* between the services received by those given access to JTPA and the services they would have received if they had been excluded from the program." They add that "the benchmark against which we measure the effects of JTPA is the services available elsewhere in the community, not a total absence of services." As illustrated by our example, this is simply a different way of saying that in the presence of substitution bias, the outcomes of the control group do not correspond to the desired counterfactual, and that the experiment provides downward-biased estimates of the effect of training on the trained. It defines the effect of JTPA relative to unspecified alternatives that vary among the 16 sites and among persons at each site.

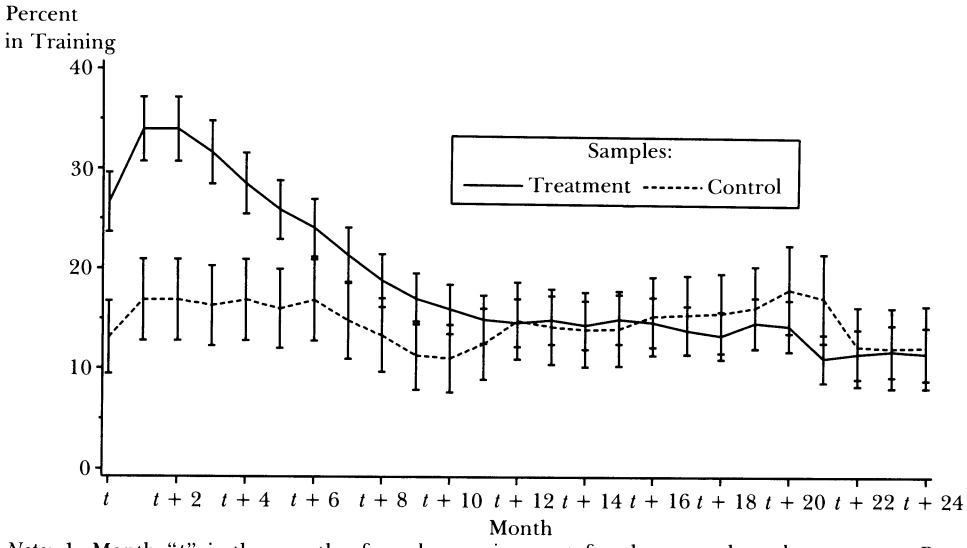
The view that the experimental estimates correspond to a counterfactual in which the JTPA program does not exist is also incorrect. The level of alternative training services available is not exogenous; it depends on the availability of the JTPA program. In its absence, both nonprofit agencies and private firms would likely increase their provision of training services.

Further dramatic evidence on substitution bias is given in Figure 2. It reports training received by young women in the JTPA experiment. Although those randomized into the control group experience a delay in receiving training, by 24 months after random assignment the same portion of the treatment and control groups had received training. JTPA is only one of many possible training options and those who wanted training managed to find it, whether or not they were randomized into the treatment group.

The problem of substitution bias is not unique to the JTPA experiment. Many previous evaluations have encountered the same problems. For example,

[11] The ensuing discussion relies on Heckman and Smith (1993, 1994b).

Figure 2
Controls and Treatments Percent in School or Training, Female Youth



Note: 1. Month "t" is the month of random assignment for the controls and treatments. Bars indicate confidence bands.
2. The monthly training information graphed here is derived from self-reported data on spells of schooling and training.

the MDRC evaluation of the Career Beginnings program noted the availability of many close substitutes (Cave and Quint, 1991, pp. 36–51). Abt's study of the Food Stamp Employment and Training Program also encountered serious problems with substitution bias (Puma et al., 1990). Persons randomized out of programs can often find good substitutes for them. An informative experimental evaluation must account for choices among the substitutes and the content of the substitutes. That is, in the presence of substitution bias it is necessary to perform a complementary nonexperimental analysis.

Summary and Conclusions

This paper has assessed the commonly made arguments concerning experimental methods of social program evaluation. While experiments can eliminate the potential for selection bias to affect mean-impact estimates, we find that the existing literature overstates many of the other arguments in their favor. There is a sizeable divergence between the theoretical capabilities of evaluations based on random assignment and the practical results of such evaluations. Moreover, experimental advocates ignore promising developments in the theory and practice of nonexperimental evaluations.

While the existing regime of self-contained black-box experimental evaluations designed to produce only mean-difference estimates of program impact supports a healthy contract research industry, it contributes next to nothing to the cumulative body of social science knowledge regarding program participation processes, earnings, wage and employment dynamics or program operation. In fact, simple black-box evaluations pose a serious threat to the accumulation of knowledge about the behavior of persons and institutions. Because they are not conducted within a behaviorally coherent framework of analysis, the evidence from experiments does not cumulate. The end result of a research program based on experiments is just a list of programs that "work" and "don't work," but no understanding of why they succeed or fail.

The long-run value of cumulative knowledge is high, but is neglected by advocates of "short-run" evaluations conducted outside of coherent social science frameworks. The potential of evaluations to add to this store of knowledge, and for this store of knowledge to inform future evaluations, needs to be more widely recognized and should be factored into current discussions regarding evaluation methodology.

■ *This research was supported by NSF grants SBR-91-11455 and SBR-93-09325, by a grant from the Lynde and Harry Bradley Foundation, Milwaukee, Wisconsin, and by a grant from the Russell Sage Foundation. We thank Alan Auerbach, Carl Shapiro, and Timothy Taylor for their thoughtful comments.*

References

Andrews, Donald, "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Regression Models," *Econometrica*, March 1991, *59*, 307–45.

Angrist, Joshua, and Guido Imbens, "Sources of Identifying Information in Evaluation Models," unpublished paper, Harvard University, 1991.

Ashenfelter, Orley, "Estimating the Effect of Training Programs on Earnings," *Review of Economics and Statistics*, February 1978, *60*, 47–57.

Ashenfelter, Orley, and David Card, "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," *Review of Economics and Statistics*, November 1985, *67*, 648–60.

Auerbach, Alan, and Larry Kotlikoff, *Dynamic Fiscal Policy*. New York: Cambridge University Press, 1987.

Bloom, H., L. Orr, G. Cave, S. Bell, and F. Doolittle, *The National JTPA Study: Title IIA Impacts on Earnings and Employment at 18 Months*. Bethesda, Md.: Abt Associates, January 1993.

Burtless, Gary, "The Case for Social Experiments." In Jensen, Karsten, and Per Kongshoj Madsen, eds., *Measuring Labour Market Measures: Evaluating the Effects of Active Labour Market Policies*. Copenhagen: Ministry of Labour, 1993, pp. 15–34.

Burtless, Gary, and Larry Orr, "Are Classical Experiments Needed for Manpower Policy?," *Journal of Human Resources*, Fall 1986, *21*, 606–39.

Cain, Glen, and Douglas Wissoker, "A Reanalysis of Marital Stability in the Seattle-

Denver Income Maintenance Experiment," *American Journal of Sociology*, March 1990, *95*, 1235–69.

Cave, George, and Janet Quint, *Career Beginnings Impact Evaluation: Findings from a Program for High School Students*. New York: Manpower Demonstration Research Corporation, 1991.

Clements, Nancy, James Heckman, and Jeffrey Smith, "Making the Most Out of Social Experiments: Reducing the Intrinsic Uncertainty in Evidence from Randomized Trials with an Application to the National JTPA Experiment," unpublished paper, University of Chicago, 1993; *Review of Economic Studies*, forthcoming.

Cosslett, Steven, "Semiparamteric Estimation of a Regression Model With Sample Selectivity." In Barnett, W. A., J. Powell, and G. Tauchen, eds., *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. Cambridge: Cambridge University Press, 1991, pp. 175–97.

Doolittle, Fred, and Linda Traeger, *Implementing the National JTPA Study*. New York: Manpower Demonstration Research Corporation, April 1990.

Fraker, Thomas, and Rebecca Maynard, "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs," *Journal of Human Resources*, Spring 1987, *22*, 194–227.

Frechet, M., "Sur Les Tableux de Correlation Dont Les Marges Sont Données," *Ann. University: Lyon Sect. A*, 1951, ser. 3, *14*, 53–77.

Hannan, Michael, and Nancy Tuma, "A Reassessment of the Effect of Income on Marital Dissolution in the Seattle-Denver Experiment," *American Journal of Sociology*, March 1990, *95*, 1270–98.

Hansen, Janet, ed., *Preparing for the Workplace: Charting a Course for Federal Post-Secondary Policy*. Washington, D. C.: National Research Council, 1994.

Hausman, Jerry and David Wise, "Technical Problems in Social Experimentation: Cost Versus Ease of Analysis." In Hausman, Jerry, and David Wise, eds., *Social Experimentation*. Chicago: University of Chicago Press for NBER, 1985, pp. 187–208.

Heckman, James, "Varieties of Selection Bias," *American Economic Review*, May 1990a, *80*, 313–18.

Heckman, James, "Alternative Approaches to the Evaluation of Social Programs: Econometric and Experimental Methods," lecture, World Congress of the Econometric Society, Barcelona, August 1990b.

Heckman, James, "Randomization and Social Program Evaluation." In Manski, Charles, and Irwin Garfinkel, eds., *Evaluating Welfare and Training Programs*. Cambridge, Mass.: Harvard University Press, 1992, pp. 201–30.

Heckman, James "Randomization as a Multiple Instrumental Variable," unpublished paper, University of Chicago, 1993a.

Heckman, James, "The Case for Simple Estimators: Experimental Evidence from the National JTPA Study," unpublished paper, University of Chicago, 1993b.

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*, December 1989, *84*:408, 862–80.

Heckman, James, and Richard Robb, "Alternative Methods for Evaluating the Impact of Interventions." In Heckman, J., and B. Singer, eds., *Longitudinal Analysis of Labor Market Data*. New York: Cambridge University Press, 1985, pp. 156–245.

Heckman, James, and Rebecca Roselius, "Evaluating the Impact of Training on the Earnings and Labor Force Status of Young Women: Better Data Help A Lot," unpublished paper, University of Chicago, 1994.

Heckman, James, and Jeffrey Smith, "Assessing the Case for Randomized Evaluation of Social Programs." In Jensen, Karsten, and Per Kongshoj Madsen, eds., *Measuring Labour Market Measures: Evaluating the Effects of Active Labour Market Policies*. Copenhagen: Ministry of Labour, 1993, pp. 35–95.

Heckman, James, and Jeffrey Smith, "Ashenfelter's Dip and the Determinants of Participation in a Social Program: Implications for Simple Program Evaluation Strategies," unpublished paper, University of Chicago, 1994a.

Heckman, James, and Jeffrey Smith, "Substitution Bias in Social Experiments: An Analysis of the JTPA Data," unpublished paper, University of Chicago, 1994b.

Heckman, James, and Jeffrey Smith, "Evaluating the Welfare State," presented at Frisch Symposium in Oslo, March 1995.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd, "Non-Parametric Characterization of Selection Bias Using Experimental Data: A Study of Adult Males in JTPA," unpublished paper, University of Chicago, 1995.

Heckman, James, Jeffrey Smith, and Christopher Taber, "Accounting for Dropouts in Evaluations of Social Experiments." NBER Technical Working Paper No. 166, 1994.

Hoeffding, W., "Masstabinvariate Korrelationstheorie," *Schriften des Mathematischen Instituts und des Institutes für Angewandte Mathematik der Universität Berlin*, 1940, 179–251.

Hotz, V. Joseph, "Designing an Evaluation of JTPA." In Manski, Charles, and Irwin Garfinkel, eds., *Evaluating Welfare and Training Programs*. Cambridge, Mass: Harvard University Press, 1992, pp. 76–114.

Hotz, V. Joseph, and Seth Sanders, "Bounding Treatment Effects in Controlled and Natural Experiments Subject to Post-Randomization Treatment Choice." Center for Social Program Evaluation Working Paper, University of Chicago, 1994.

Ichimura, Hidehiko, and Lung-Fei Lee, "Semiparametric Least Squares Estimation of Multiple Index Models: Single Equation Estimation." In Barnett, W. A., J. Powell, and G. Tauchen, eds., *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. Cambridge: Cambridge University Press, 1991, pp. 3–49.

Kramer, Michael, and Stanley Shapiro, "Scientific Challenges in the Application of Randomized Trials," *Journal of the American Medical Association*, November 16, 1984, *252*, 2739–45.

Kydland, Finn, and Edward Prescott, "The Econometrics of the General Equilibrium Approach to Business Cycles," *Scandinavian Journal of Economics*, 1991, *93*:2, 161–78.

LaLonde, Robert, "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, September 1986, *76*, 604–20.

Moffitt, Robert, "Evaluation Methods for Program Entry Effects." In Manski, Charles, and Irwin Garfinkel, eds., *Evaluating Welfare and Training Programs*. Cambridge, Mass: Harvard University Press, 1992, pp. 231–52.

Newey, Whitney, "Two Step Series Estimation of Sample Selection Models," unpublished paper, Princeton University, 1988.

Orcutt, Guy, and Alice Orcutt, "Experiments For Income Maintenance Policies," *American Economic Review*, September 1968, *58*:4, 754–72.

Powell, James, "Semiparametric Estimation of Censored Selection Models," unpublished paper, University of Wisconsin-Madison, 1989.

Puma, Michael, N. Burstein, K. Merrell, and G. Silverstein, *Evaluation of the Food Stamp Employment and Training Program: Final Report*. Bethesda, Md.: Abt Associates, June 1990.

Shoven, John, and John Whalley, *Canada-U.S. Tax Comparisons*. Chicago: University of Chicago Press for NBER, 1992.

Tinbergen, Jan, *Economic Policy: Principles and Design*. Amsterdam:North-Holland, 1956.

Torp, Hege, O. Rauum, E. Hernaes, and H. Goldstein, "The First Norwegian Experiment." In Jensen, Karsten, and Per Kongshoj Madsen, eds., *Measuring Labour Market Measures: Evaluating the Effects of Active Labour Market Policies*. Copenhagen: Ministry of Labour, 1993, pp. 97–140.

Zellner, Arnold, and Peter Rossi, "Evaluating the Methodology of Social Experiments." In Munnell, Alicia, ed., *Lessons from the Income Maintenance Experiments*. Boston: Federal Reserve Bank of Boston, 1986, pp. 131–57.