Barnow, Burt S.

Article

Setting up social experiments: the good, the bad, and the ugly

Zeitschrift für ArbeitsmarktForschung - Journal for Labour Market Research

Provided in Cooperation with:
Institute for Employment Research (IAB)

Suggested Citation: Barnow, Burt S. (2010) : Setting up social experiments: the good, the bad, and the ugly, Zeitschrift für ArbeitsmarktForschung - Journal for Labour Market Research, ISSN 2510-5027, Springer, Heidelberg, Vol. 43, Iss. 2, pp. 91-105, https://doi.org/10.1007/s12651-010-0042-6

This Version is available at:
https://hdl.handle.net/10419/158720

ZAF (2010) 43:91–105
DOI 10.1007/s12651-010-0042-6

INVITED PAPER

Setting up social experiments: the good, the bad, and the ugly

Burt S. Barnow

Published online: 20 October 2010

Abstract It is widely agreed that randomized controlled trials – social experiments – are the gold standard for evaluating social programs. There are, however, many important issues that cannot be tested using social experiments, and often things go wrong when conducting social experiments. This paper explores these issues and offers suggestions on ways to deal with commonly encountered problems. Social experiments are preferred because random assignment assures that any differences between the treatment and control groups are due to the intervention and not some other factor; also, the results of social experiments are more easily explained and accepted by policy officials. Experimental evaluations often lack external validity and cannot control for entry effects, scale and general equilibrium effects, and aspects of the intervention that were not randomly assigned. Experiments can also lead to biased impact estimates if the control group changes its behavior or if changing the number selected changes the impact. Other problems with conducting social experiments include increased time and cost, and legal and ethical issues related to excluding people from the treatment. Things that sometimes go wrong in social experiments include programs cheating on random assignment, and participants and/or staff not understanding the intervention rules. The random assignment evaluation of the Job Training Partnership Act in the United States is used as a case study to illustrate the issues.

B. S. Barnow (✉)
George Washington University,
Trachtenberg School of Public Policy and Public Administration,
805 21st St, NW, Washington, DC 20052, USA

Die Gestaltung von Sozialexperimenten: The good, the bad and the ugly

Zusammenfassung Es herrscht weitestgehend Konsens darüber, dass randomisierte kontrollierte Studien – Sozialexperimente – der „Goldstandard" für die Bewertung sozialer Programme sind. Es gibt jedoch viele wichtige Aspekte, die sich nicht durch solche Studien bewerten lassen, und bei der Durchführung dieser Studien kann oft etwas schiefgehen. Die vorliegende Arbeit untersucht diese Themen und bietet Lösungsvorschläge für häufig auftretende Probleme. Sozialexperimente werden bevorzugt, weil die Randomisierung dafür sorgt, dass alle Unterschiede zwischen der Treatmentgruppe und der Kontrollgruppe der Intervention und nicht einem anderen Faktor zuzuschreiben sind. Es fällt Politikern und Beamten auch leichter, die Ergebnisse von Sozialexperimenten zu erklären und zu akzeptieren.

Bei experimentellen Bewertungen fehlt oft die externe Validität, und es fehlt die Möglichkeit, „entry effects", Skaleneffekte, allgemeine Gleichgewichtseffekte und nichtrandomisierte Aspekte der Intervention zu kontrollieren. Experimente können auch zu verzerrten Aussagen über die Auswirkungen führen, wenn die Kontrollgruppe ihr Verhalten ändert oder wenn eine Änderung der Anzahl der ausgewählten Personen zu einer Veränderung der Auswirkungen führt. Weitere Probleme bei Sozialexperimenten sind erhöhter Zeitaufwand und Kosten sowie juristische und ethische Fragen nach dem Ausschluss gewisser Menschen von den Maßnahmen. Fehler, die manchmal in Sozialexperimenten vorkommen, sind beispielsweise Programme, die bei der Randomisierung nicht korrekt vorgehen und Teilnehmer bzw. Mitarbeiter, die die Interventionsregeln nicht verstehen. Die randomisierte Bewertung des Job Training Partnership Act in den USA wird als Fallstudie verwendet, um diese Themen besser aufzuzeigen.

1 Introduction

Since the 1960s, social experiments have been increasingly used in the United States to determine the effects of pilots and demonstrations as well as ongoing programs in areas as diverse as education, health insurance, housing, job training, welfare cash assistance, and time of day pricing of electricity. Although social experiments have not been widely used in Europe, there is growing interest in expanding their use in evaluating social programs. Social experiments remain popular in the United States, but there has been a spirited debate in recent years regarding whether recent methodological developments, particularly propensity score matching and regression discontinuity designs, overcome many of the key objections to nonexperimental methods. This paper provides an assessment of some of the issues that arise in conducting social experiments and explains some of the things that can go wrong in conducting and interpreting the results of social experiments.

The paper first defines what is generally meant by the term social experiments and briefly reviews their use in the United States. This is followed by a discussion of the advantages of social experiments over nonexperimental methods. The next section discusses the limitations of social experiments – what we *cannot* learn from social experiments. Next is a section discussing some of the things that can go wrong in social experiments and limits of what we learn from them. To illustrate the problems that can arise, the penultimate section provides a case study of lessons from the National JTPA Study, a social experiment that was used to assess a large training program for disadvantaged youth and adults in the United States. The last section provides conclusions.

2 Definitions and context

As Orr (1999, p. 14) notes, "The defining element of a social experiment is random assignment of some pool of individuals to two or more groups that are subject to different policy regimes." Greenberg and Shroder (2004, p. 4) note that because social experiments are intended to provide unbiased estimates of the impacts of the policy of interest, they must have four specific features:

- *Random assignment*: Creation of at least two groups of human subjects who differ from one another by chance alone.
- *Policy intervention*: A set of actions ensuring that different incentives, opportunities, or constraints confront the members of each of the randomly assigned groups in their daily lives.
- *Follow-up data collection*: Measurement of market and fiscal outcomes for members of each group.
- *Evaluation*: Application of statistical inference and informed professional judgment about the degree to which the policy interventions have caused differences in outcomes between the groups.

These four features are not particularly restrictive, and social experiments can have a large number of variations. Although we often think of random assignment taking place at the individual level, the random assignment can take place at a more aggregated level, such as the classroom, the school, the school district, political or geographic jurisdictions, or any other unit where random assignment can be feasibly carried out.[1] Second, there is no necessity for a treatment to be compared against a null treatment. In an educational or medical context, for example, it might be harmful to the control group if they receive no intervention; in such instances, the experiment can measure differential impacts where the treatment and control groups both receive treatments, but they do not receive the same treatment.[2]

Third, there does not have to be a single treatment. In many instances it is sensible to develop a number of alternative treatments to which participants are assigned. In health insurance experiments, for example, there are often a number of variations we would like to test for the key aspects of the treatment. Thus, we might want to randomly assign participants to various combinations of deductable amounts and co-payment rates to see which combination leads to the best results in terms of costs and health outcomes. Likewise, in U.S. welfare experiments, the experiments frequently vary the "guarantee," the payment received if the person does no market work, and the "implicit tax rate," the rate at which benefits are reduced if there are earnings.[3]

Fourth, social experiments can be implemented in conjunction with an ongoing program or to test a new intervention; in some instances a social experiment will test a new intervention in the context of an ongoing program.

[1] There are a number of factors that help determine the units used for random assignment. Assignment at the individual level generates the most observations, and hence the most precision, but in many settings it is not practical to conduct random assignment at the individual level. For example, in an educational setting, it is generally not feasible to assign students in the same classroom to different treatments. The most important problem resulting from random assignment at a more aggregated level is that there are fewer observations, leading to a greater probability that the treatment and control groups are not well matched and the potential for imprecise estimates of the treatment effect.

[2] It is important to distinguish between a known null treatment and a broader "whatever they would normally get" control treatment. As discussed below, the latter situation often makes it difficult to know what comparison is specifically being made and how estimated the impacts should be interpreted.

[3] Orr (1999) notes that by including a variety of treatment doses, we can learn more than the effect of a single dose level on participants; instead, we can estimate a behavioral response function that provides information on how the impact varies with the dosage. Heckman (2008) provides a broader look at the concept of economic causality.

Welfare programs in the United States have been subject to several types of social experiments. In the 1960s and 1970s, a series of "negative income tax" experiments were conducted where a randomly selected group of people were diverted from regular welfare programs to entirely new welfare programs with quite different rules and benefits. During the 1980s and 1990s, many states received waivers where they were permitted to try new variations on their welfare programs so long as the new interventions were evaluated using random assignment. U.S. vocational training programs have included freestanding demonstrations with experimental designs as well as experimental evaluations of ongoing programs. Inserting an experimental design in an ongoing program is sometimes difficult, particularly if the program is an entitlement or if the authorizing legislation prohibits denying services to those who apply.

Another important distinction among experiments is that the participants can volunteer for the intervention or they can be assigned to the program. For purely voluntary programs, such as many job training programs in the United States, there is no meaningful concept of mandatory participants. For welfare programs, however, a new intervention can be voluntary in nature or it could be mandatory; the numerous welfare to work demonstration programs tested in the United States have fallen into both categories. While both mandatory and voluntary programs can be evaluated using an experimental design, the findings must be interpreted carefully. The impacts estimated for a voluntary program can not necessarily be expected to apply for a program where all welfare recipients must participate, and the impacts for a mandatory program may not apply if the same intervention were implemented as a voluntary program.

Although this paper does not focus on the ethics of random assignment, it is important to consider whether it is ethical to deny people the opportunity to participate in a social program. Both Greenberg and Shroder (2004) and Orr (1999) discuss the ethics of random assignment, but they do not do so in depth. More recently, the topic was explored in more depth in an exchange between Blustein (2005a,b), Barnow (2005), Rolston (2005), and Schochet (2005). Many observers would agree that random assignment is ethical (or at least not unethical) when there is excess demand for a program and the effectiveness of the program is unknown. Blustein (2005a) uses the experimental evaluation of the Job Corps to raise issues such as recruiting additional applicants so that there will be sufficient applicants to deny services to some, the fact that applicants who do not consent to the random assignment procedure are denied access to the program, and whether those randomized out of participation should receive monetary compensation. She believes that a good case can be made that the Job Corps evaluation, which included random assignment, may have been unethical, although her critics generally take issue with her points and claim that the knowledge gained is sufficient to offset any losses to the participants. As Blustein makes clear, her primary motivation in the paper is not to dispute the ethics of the Job Corps evaluation but rather to urge that ethical considerations be taken into account more fully when random assignment is being considered.

An important distinction between social experiments and randomized controlled trials that are frequently used in the fields of medicine and public health is that social experiments rarely make use of double blind or even single blind approaches. In the field of medicine, it is well known that there can often be a "placebo effect" where subjects benefit from the perception of such a treatment. Although social experiments can also be subject to similar problems, it is often difficult or impossible to keep the subjects and researchers unaware of their treatment status. A related phenomenon, known as the "Hawthorne effect," refers to the possibility that subjects respond differently to stimuli because they are being observed.[4] The important point is that the inability to conduct double blind experiments, and even the knowledge that a subject is in an experiment can potentially lead to biased estimates of intervention impacts.

It is important to distinguish between true social experiments and "natural experiments." The term natural experiment is sometimes used to refer to situations where random selection is not used to determine assignment to treatment status but the mechanism used, it is argued, results in treatment and comparison groups that are virtually identical. Angrist and Krueger (2001) extol the use of natural experiments in evaluations when random assignment is not feasible as a way to eliminate omitted variable bias; however, the examples they cite make use of instrumental variables rather than assuming that simple analysis of variance or ordinary least squares regression analysis can be used to obtain impact estimates:

> Instruments that are used to overcome omitted variable bias are sometimes said to derive from "natural experiments." Recent years have seen a resurgence in the use of instrumental variables in this way – that is, to exploit situations where the forces of nature or government policy have conspired to produce an environment somewhat akin to a randomized experiment. This type of application has generated some of the most provocative empirical findings in economics, along with some controversy over substance and methods.

Perhaps one of the best known examples of use of a natural experiment is the analysis by Angrist and Krueger (1991) to evaluate the effects of compulsory school attendance

[4] There are many views on how serious Hawthorne effects distort impact estimates, in the original illumination studies at the Hawthorne works in the 1930s and in other contexts.

laws in the United States on education and earnings. In that study, the authors argue that the number of years of compulsory education (within limits) is essentially random, as it is determined by the month of birth. As Angrist and Krueger clearly imply, a natural experiment is not a classical experiment with randomized control trials, and there is no guarantee that simple analyses or more complex approaches such as instrumental variables will yield unbiased treatment estimates.

3 Why conduct social experiments?

There are a number of reasons why social experiments are preferable to nonexperimental evaluations. In the simplest terms, the objective in an evaluation of a social program is to observe the outcome for an intervention for the participants with and without the intervention. Because it is impossible to observe the same person in two states of the world at the same time, we must rely on some alternative approach to estimate what would have happened to participants had they not been in the program. The simplest and most effective way to assure comparability of the treatment and control groups is to randomly assign the potential participants to either receive the treatment or be denied the treatment; with a sufficiently large sample size, the treatment and control groups are likely to be identical on all characteristics that might affect the outcome. Nonexperimental evaluation approaches generally seek to provide unbiased and consistent impact estimates either by using mechanisms to develop comparison groups that are as similar as possible to the treatment group (e.g., propensity score matching) or by using econometric approaches to control for observed and unobserved omitted variables (e.g., fixed effects models, instrumental variables, ordinary least squares regression analysis, and regression discontinuity designs). Unfortunately, all the nonexperimental approaches require strong assumptions to assure that unbiased estimates are obtained, and these assumptions are not always testable.

Burtless (1995) describes four reasons why experimental designs are preferable to nonexperimental designs. First, random assignment assures the direction of causality. If earnings rise for the treatment group in a training program more than they do for the control group, there is no logical source of the increase other than the program. If a comparison group of individuals who chose not to enroll is used, the causality is not clear – those who enroll may be more interested in working and it is the motivation that leads to the earnings gain rather than the treatment. Burtless's second argument is related to the first – random assignment assures that there is no selection bias in the evaluation, where selection bias is defined as a likelihood that individuals with particular unobserved characteristics may be more or less likely to participate in the program.[5] The most common example of potential selection bias is that years of educational attainment are likely to be determined in part on ability, but ability is usually either not available to the evaluator or available only with measurement error.

The third argument raised by Burtless in favor of social experiments is that social experiments permit tests of interventions that do not naturally occur. Although social experiments do permit evaluations of such interventions, pilot projects and demonstrations can also be implemented without a randomly selected control group. Finally, Burtless notes that evaluations using random assignment provide findings that are more persuasive to policy makers than evaluations using nonexperimental methods. One of the best features of using random assignment is that program impacts can be observed by simply subtracting the post-program control group values from the values for the treatment group – there is no need to have faith that a fancy instrumental variables approach or a propensity score matching scheme has adequately controlled for all unobserved variables.[6] For researchers, experiments also assure that the estimates are unbiased and more precise than alternative approaches.

4 Can nonexperimental methods replicate experimental findings?

The jury is still out on this issue, and in recent years there has been a great deal of research and spirited debate about how well nonexperimental methods do at replicating experimental findings, given the data that are available. There is no question that there have been important developments in nonexperimental methods in recent years, but the question remains as to how well the methods do in replicating experimental findings and how the replication depends on the particular methods used and data available. Major contributions in recent years include the work of Heckman et al. (1997) on propensity score matching and Hahn et al. (2001) on regression discontinuity designs.[7] In this section several recent studies that have found a good match between nonexperimental methods and experimental findings are first re-

[5] See Barnow et al. (1980) for a discussion of selection bias and a summary of approaches to deal with the problem.

[6] As discussed more in the sections below, many circumstances can arise that make experimental findings difficult to interpret.

[7] Propensity score matching is a two-step procedure where in the first stage the probability of participating in the program is estimated, and, in the simplest approach, in the second stage the comparison group is selected by matching each member of the treatment group with the nonparticipating person with the closest propensity score; there are numerous variations involving techniques such as multiple matches, weighting, and calipers. Regression discontinuity designs involve selection mechanisms where treatment/control status is determined by a screening variable.

viewed, followed by a review of studies that were unable to replicate experimental findings. The section concludes with suggestions from the literature on conditions where nonexperimental approaches are most likely to replicate experimental findings.

Propensity score matching has been widely used in recent years when random assignment is not feasible. Heckman et al. (1997) tested a variety of propensity score matching approaches to see what approaches best mirror the experimental findings from the evaluation of the Job Training Partnership Act (JTPA) in the United States. The authors conclude that: "We determine that a regression-adjusted semiparametric conditional difference in differences matching estimator often performs the best among a class of estimators we examine, especially when omitted time-invariant characteristics are a source of bias." The authors caution, however: "As is true of any empirical study, our findings may not generalize beyond our data." They go on to state: "Thus, it is likely that the insights gained from our study of the JTPA programme on the effectiveness of different estimators also apply in evaluating other training programmes targeted toward disadvantaged workers."

Another effort to see how well propensity score matching replicates experimental findings is in Dehejia and Wahba (2002). These authors are also optimistic about the capability of propensity score matching to replicate experimental impact estimates: "This paper has presented a propensity score-matching method that is able to yield accurate estimates of the treatment effect in nonexperimental settings in which the treated group differs substantially from the pool of potential comparison units." Dehejia and Wahba (2002) use propensity score matching in trying to replicate the findings from the National Supported Work demonstration. Although the authors find that propensity score matching works well in the instance they examined, they caution that the approach critically depends on selection being based on observable variables and note that the approach may not work well when important explanatory variables are missing.

Cook et al. (2008) provide a third example of finding that nonexperimental approaches do a satisfactory job of replicating experimental findings under some circumstances. The authors looked at the studies by the type of nonexperimental approach that was used. The three studies that used a regression discontinuity design were all found to replicate the findings from the experiment.[8] They note that although regression discontinuity designs are much less efficient than experiments, as shown by Goldberger (1972), the studies they reviewed had large samples so impacts remained statistically significant. The authors find that propensity score matching works well in replicating experimental findings when key covariates are included in the propensity score modeling and where the comparison pool members come from the same geographic area as the treatment group, and they also find that propensity score matching works well when clear rules for selection into the treatment group are used and the variables that are used in selection are available for the analysis. Finally, in studies where propensity score matching was used but the covariates available did not correspond well to the selection rules and/or there was a poor geographic match, the nonexperimental results did not consistently match the experimental findings.

In another recent study, Shadish et al. (2008) conducted an intriguing experiment by randomly assigning one group of individuals to be randomly assigned to treatment status and the other to self-select one of the two treatment options (mathematics or vocabulary training). The authors found that propensity score matching greatly reduced the bias of impact estimates when the full set of available covariates was used, including pretests, but did poorly when only predictors of convenience (sex, age, marital status, and ethnicity) were used. Thus, their findings correspond with the findings of Cook et al. (2008).

Smith and Todd (2005a) reanalyzed the National Supported Work data used by Dehejia and Wahba (2002). They find that the estimated impacts are highly sensitive to the particular subset of the data analyzed and the variables used in the analysis. Of the various analytical strategies employed, Smith and Todd (2005a) find that difference in difference matching estimators perform the best. Like many other researchers, Smith and Todd (2005a) find that variations in the matching procedure (e.g., number of individuals matched, use of calipers, local linear regressions) generally do not have a large effect on the estimated impacts. Although they conclude that propensity score matching can be a useful approach for nonexperimental evaluations, they believe that it is not a panacea and that there is no single best approach to propensity score matching that should be used.[9]

Wilde and Hollister (2007) used data from an experimental evaluation of a class size reduction effort in Tennessee (Project STAR) to assess how well propensity score matching replicates the experimental impact estimates. They accomplished this by treating each school as a separate experiment and pooling the control groups from other schools in the study and then using propensity score matching to identify the best match for the treatment group in each school. The authors state that: "Our conclusion is that propensity

[8] It is important to keep in mind that regression discontinuity designs provide estimates of impact near the discontinuity, but experiments provide estimates over a broader range of the population.

[9] See also the reply by Dehejia (2005) and the rejoinder by Smith and Todd (2005b).

score estimators do not perform very well, when judged by standards of how close they are to the 'true' impacts estimated from experimental estimators based on a random assignment design."[10]

Bloom et al. (2002) make use of an experiment designed to assess the effects of mandatory welfare to work programs in six states to compare a series of comparison groups and estimation strategies to see if popular nonexperimental methods do a reasonable job of approximating the impact estimates obtained from the experimental design. Nonexperimental estimation strategies tested include several propensity score matching strategies, ordinary least squares regression analysis, fixed effect models, and random growth models. The authors conclude that none of the approaches tried do a good job of reproducing the experimental findings and that more sophisticated approaches are sometimes worse than simple approaches such as ordinary least squares.

Overall, the weight of the evidence appears to indicate that nonexperimental approaches generally do not do a good job of replicating experimental estimates and that the most common problem is the lack of suitable data to control for key differences between the treatment group and comparison group. The most promising nonexperimental approach appears to be the regression discontinuity design, but this approach requires a much larger sample size to obtain the same amount of precision as an experiment.[11] The studies identify a number of factors that generally improve the performance of propensity score matching:

- It is important to only include observations in the region of common support, where the probabilities of participating are nonzero for both treatment group members and comparison group members.
- Data for the treatment and comparison groups should be drawn from the same data source, or the same questions should be asked of both groups.
- Comparison group members should be drawn from the same geographic area as the treatment group.
- It is important to understand and statistically control for the variables used to select people into the treatment group and to control for variables correlated with the outcomes of interest.
- Difference in difference estimators appear to produce less bias than cross section matching in several of the studies, but it is not clear that this is always the case.

[10] The paper by Wilde and Hollister (2007) is one of the papers reviewed by Cook et al. (2008), and they claim that because Wilde and Hollister control on too few covariates and draw their comparison group from other areas than where the treatment group resides, the Wilde and Hollister paper does not offer a good test of propensity score matching.

[11] Schochet (2009) shows that a regression discontinuity design typically requires a sample three to four times as large as an experimental design to achieve the same level of statistical precision.

5 What we cannot learn from social experiments

Although experiments provide the best means of obtaining unbiased estimates of program impacts, there are some important limitations that must be kept in mind in designing experiments and interpreting the findings. This section describes some of the limitations that are typically inherent to experiments as well as problems that sometimes arise in experiments.

Although a well designed experiment can eliminate internal validity problems, there are often issues regarding external validity, the applicability of the findings in other situations. External validity for the eligible population is threatened if either the participating sites or individuals volunteer for the program rather than are randomly assigned. If the sites included in the experiment volunteered rather than were randomly selected, the impact findings may not be applicable to other sites. It is possible that the sites that volunteer are more effective sites, as less capable sites may want to avoid having their poor performance known to the world. In some of the welfare to work experiments conducted in the United States, random assignment was conducted among welfare recipients who volunteered to participate in the new program. The fact that the experiment was limited to welfare recipients who volunteered would not harm the internal validity of the evaluation, but the results might not apply to individuals who did not volunteer. If consideration is being given to making the intervention mandatory, then learning the effects of the program for volunteers does not identify the parameter of interest unless the program has the same impact on all participants. Although there is no way to assure external validity, exploratory analyses examining whether impacts are consistent across sites and subgroups can suggest (but not prove) if there is a problem.

Experiments typically randomly assign people to the treatment or control group *after* they have applied for or enrolled in the program. Thus, experiments typically do not pick up any effects the intervention might have that encourage or discourage participation. For example, if a very generous training option is added to a welfare program, more people might sign up for the program. These types of effects, referred to as entry effects, can be an important aspect of a program's effects. Because experiments are likely not to measure these effects, nonexperimental methods must be used to estimate the entry effects.[12]

Another issue that is difficult to deal with in the context of experiments is the finite time horizon that typically accompanies an experiment. If the experiment is offered on a temporary basis and potential participants are aware of

[12] See Moffitt (1992) for a review of the topic and Card and Robins (2005) for a recent evaluation of entry effects.

the finite period of the experiment, their behavior may be quite different from what would occur if the program were permanent. Consider a health insurance experiment, for example. If members of the treatment group have more generous coverage during the experiment than they will have after the experiment, they are more likely to increase their spending on health care for services that might otherwise be postponed. The experiment will provide estimates of the impact of a temporary policy, but what is needed for policy purposes is the impact of a permanent program. This issue can be dealt with in several ways. One approach would be to run the experiment for a long time so that the treatment group's response would be similar to what would occur for a permanent program; this would usually not be feasible due to cost issues. Another approach would be to enroll members of the treatment group for a varying number of years and then try to estimate how the response varies with time in the experiment. Finally, one could enroll the participants in a "permanent" program and then buy them out after the data for the evaluation has been gathered.

Another area where experiments may provide only limited information is on general equilibrium effects. For example, a labor market intervention can have effects not captured in a typical evaluation. Examples include potential displacement of other workers by those who receive training, wage increases for the control group due to movement of those trained into a different labor market, and negative wage effects for occupations if the number of people trained is large. Another example is "herd immunity" observed in immunization programs; the benefits of an immunization program affect those not immunized at some point as their probability of contracting the disease diminishes as the number of people in the community immunized increases. Not only do small scale experiments fail to measure these effects, even the evaluation of a large scale program might miss them.[13]

With human subjects, it is not always a simple matter to assure that individuals in the treatment group obtain the treatment and those in the control group do not receive the treatment. In addition, being in the control group in the experiment may provide benefits that would not have been received had there been no experiment. These three cases are described below.

One factor that differentiates social experiments from agricultural experiments is that often some of those assigned to the treatment group do not receive the treatment. So-called no-shows are frequently found in program evaluations, including experiments. It is essential that no-shows be included in the treatment group to preserve the equality of the treatment and control groups. Unfortunately, the experimental impact estimates produced when there are no-shows provide the impact of an offer of the treatment, not the impact of the treatment itself. A policy maker who is trying to decide whether to continue a training program is not interested in the impact of an offer for training – the program only incurs costs for those who enroll, so the policy maker wants to know the impact for those who participate.

Bloom (1984) has shown that if one is willing to assume that the treatment has no impact on no-shows, the experimental impact estimator can be adjusted to provide an estimate of the impact on the treated. The overall impact of the program is a weighted average of the impact on those who receive the treatment, I_{P}, and those who do not receive the treatment, I_{NP}:

$$I = pI_{\mathrm{P}} + (1-p)I_{\mathrm{NP}}\,,$$

where p is the fraction of the treatment group that receives the treatment. If the impact on those who do not receive the treatment is zero, then $I_{\mathrm{NP}} = 0$, and $I_{\mathrm{P}} = I/p$; in other words, the impact of the program on those who receive the treatment is estimated by dividing the impact on the overall treatment group (including no-shows) by the proportion who actually receive the treatment.

Individuals assigned to the control group who somehow receive the treatment are referred to as "crossovers." Orr (1999) observes that some analysts assign the crossovers to the treatment group or leave them out of the analysis, but either of these strategies is likely to destroy the similarity of the treatment and control groups. He further observes that if we are willing to assume that the program is equally effective for the crossovers and the "crossover-like" individuals in the treatment group, then the impact on the crossover-like individuals is zero and the overall impact of the program can be expressed as a weighted average of the impact on the crossover-like individuals and other individuals:

$$I = cI_{\mathrm{c}} + (1-c)I_{\mathrm{o}}\,,$$

where I_{c} is the impact on crossover-like participants, I_{o} is the impact on others, and c is the proportion of the control group that crossed over; assuming that $I_{\mathrm{c}} = 0$, we can then compute the impact on those who do not cross over as $I_{\mathrm{o}} = I/(1-c)$. If the crossovers receive a similar but not identical treatment, then the impact on the crossover-like individuals may well not be zero, and Orr (1999) indicates that the best that can be done is to vary the value of I_{c} and obtain a range of estimates.[14]

[13] See Lise et al. (2005) for further discussion of these issues.

[14] See Heckman et al. (2000) for discussion of this issue and estimates for JTPA. The authors find that JTPA provides only a small increase in the opportunity to receive training and that both JTPA and its substitutes increase earnings for participants; thus, focusing only on the experimental estimates of training impacts can lead to a large underestimate of the impact of training on earnings.

Heckman and Smith (1995) raise a related issue. In some experiments, the control group may receive valuable services in the process of being randomized out that they would not receive if there were no experiment. This may occur because when people are being recruited for the experiment, they receive some services with the goal of increasing their interest. Alternatively, to reduce ethical concerns, those randomized out may receive information about alternative treatments, which they then receive. In either case, the presence of the experiment has altered the services received by the control group and this creates what Heckman and Smith (1995) refer to as "substitution bias."

Heckman and Smith (1995) also discuss the concept of "randomization bias" that can arise because the experiment changes the scale of the intervention. This problem can arise when the program has heterogeneous impacts and as the scale of the program is increased, those with smaller expected impacts are more likely to enroll. Suppose, for example, that at its usual scale a training program has an earnings impact of $1,000 per year. When the experiment is introduced, the number of people accepted into the program increases, so the impact is likely to decline. It is possible, at least in theory, to assess this problem and correct for it by asking programs to indicate which individuals would have been accepted at the original scale and at the experiment scale. Another possible way to avoid this problem is to reduce the operating scale of the program during the experiment so that the size of the treatment and control groups combined is equal to the normal operating size of the program. More practically, randomization bias can be minimized if the proportion randomized out is very small, say 10% or less; this was the strategy employed in the experimental evaluation of the Job Corps in the United States where Schochet (2001) indicates that only about 7% of those admitted to the program were assigned to the control group.[15]

6 What can go wrong in social experiments?

In addition to the issues described above that frequently arise in social experiments, there are a number of problems that can also arise. Several common problems are described in this section, and the following section provides a case study of one experiment.

For demonstration projects and for new programs, the intervention may change after the program is initiated. In some cases it may take several months for the program to be working at full capacity; those who enroll when the program first opens may not receive the same services as later participants receive. The program might also change because program officials learn that some program components do not work as well in practice as they do in theory, economic conditions change, or the participants differ from what was anticipated. Some types of interventions, such as comprehensive community initiatives are expected to change over their implementation as new information is gathered.[16] Although program modifications often improve the intervention, they can complicate the evaluation in several ways. Instead of determining the impact of one known intervention, the impact evaluation may provide estimates that represent an average of two or more different strategies. At worst, policy makers might believe that the impact findings apply to a different intervention than what was evaluated.

Several strategies can be used to avoid or minimize these types of problems. First, it is important to monitor the implementation of the intervention. Even ongoing programs should be subject to implementation studies so that policy makers know what is being evaluated and if it has changed over time. Second, for a new intervention, it is often wise to postpone the impact evaluation until the intervention has achieved a steady state. Finally, if major changes in the intervention occur over the period analyzed, the evaluation can be conducted for two or more separate periods, although this strategy reduces the precision of the impact estimates.

Experiments can vary in their complexity, and this can lead to problems in implementation and the interpretation of findings. In some instances, experiments are complex because we wish to determine an entire "response surface" rather than evaluate a single intervention. Examples in the United States include the RAND health insurance experiment and the negative income tax (welfare reform) experiments (Greenberg and Schroder 2004), where various groups in the experiment were subject to variations in key parameters. For example, in the negative income tax experiments, participants were subject to variation in the maximum benefit and the rate at which benefits were reduced if they earned additional income. If the participants did not understand the concepts involved, particularly the implicit tax rate on earnings, then it would be inappropriate to develop a response surface based on variation in behavior by participants subject to different rules.

Problems in understanding the rules of the intervention can also arise in simpler experiments. For example, the

[15] The Job Corps evaluation was able to deny services to a small proportion of applicants by including all eligible Job Corps applicants in the study, with only a relatively small proportion of the treatment group interviewed. The reason that this type of design has not been more actively used is that if there is a substantial fixed cost per site included in the experiment, including all sites generates large costs and for a fixed budget results in a smaller overall sample.

[16] Comprehensive community initiatives are generally complex interventions that include interventions in a number of areas including employment, education, health, and community organization. See Connell and Kubisch (1998) for a discussion of comprehensive community initiatives and why they are difficult to evaluate.

State of Maryland wished to promote good parenting among its welfare recipients and instituted an experiment called the Primary Prevention Initiative (PPI). The treatment group in this experiment was required to assure that the children in the household maintained satisfactory school attendance (80% attendance), and preschool children were required to receive immunizations and physical examinations (Wilson et al. 1999). Parents who failed to meet these criteria were subject to a fine of $25.00 per month. The experiment included an implementation study, and as part of the implementation study, clients were surveyed on their knowledge of the PPI. Wilson et al. (1999) report that "only a small minority of clients (under 20%) could correctly identify even the general areas in which PPI had behavioral requirements." The lack of knowledge was almost as high among those sanctioned as for clients not sanctioned. Not surprisingly, the impact evaluation indicated that the PPI had no effect on the number of children that were immunized, that received a physical exam, or that had satisfactory school attendance. If there had been no data on program knowledge, readers of the impact evaluation might logically have inferred that the incentives were not strong enough rather than that participants did not understand the intervention.

The potential for participants in experiments to not fully understand the rules of the intervention is not trivial. If we obtain zero impacts because participants do not understand the rules and it is possible to educate them, it is important to identify the reasons why we estimate no impact. Thus, whenever there is a reasonable possibility of participants misunderstanding the rules, it is advisable to consider including a survey of intervention knowledge as part of the evaluation.

Finally, in instances where state or local programs are asked to volunteer to participate in the program, there may be a high refusal rate, thus jeopardizing external validity. Sites with low impacts may be reluctant to participate as may sites that are having trouble recruiting adequate participants. Sites may also be reluctant to participate if they believe random assignment is unethical, as was discussed above, or adds a delay in processing applicants.

7 Lessons from the National JTPA Study

This section describes some of the problems that occurred in implementing the National JTPA Study in the United States. The Job Training Partnership Act (JTPA) was the primary workforce program for disadvantaged youth and adults in the United States from 1982 through 1998 when the Workforce Investment Act (WIA) was enacted. The U.S. Department of Labor decided to evaluate JTPA with a classical experiment after a series of impact evaluations of JTPA's predecessor produced such a wide range of estimated impacts that it was impossible to know the impact of the program.[17] The National JTPA Study used a classical experimental design to estimate the impact of the JTPA program on disadvantaged adults and out-of-school disadvantaged youth. The study began in 1986 and made use of JTPA applicants in 16 sites across the country. The impact evaluation found that the program increased earnings of adult men and women by over $1,300 in 1998 dollars during the second year after training. The study found that the out-of-school youth programs were ineffective, and these findings are not discussed below.

I focus on the interim report of the National JTPA Study for several reasons.[18] First, the study was generally well done, and it was cited by Hollister (2008) as one of the best social experiments that was conducted. The problems that I review below are not technical flaws in the study design or implementation, but program features that precluded analyzing the hypotheses of most interest and, in my view, approaches to presenting the findings that may have led policy makers to misinterpret the findings. I focus on the interim report rather than the final report because many of the presentation issues that I discuss were not repeated in the final report.[19]

7.1 Nonrandom site selection

The study design originally called for 16 to 20 local sites to be selected at random. Sites were offered modest payments to compensate for extra costs incurred and to pay for inconvenience experienced. The experiment took place when the economy was relatively strong, and many local programs (called service delivery areas or SDAs) were having difficulty spending all their funding. Because participating sites were required to recruit 50% more potential participants to construct a control group one-half the size of the treatment group, many sites were reluctant to participate in the experiment. In the end, the project enrolled all 16 sites identified that were willing and able to participate. All evaluations, including experiments, run the risk of failing to have external validity, but the fact that most local sites refused to participate raised suspicion that the sites selected did not constitute a representative sample of sites. The National JTPA Study report does note that no large cities are included in the par-

[17] See Barnow (1987) for a summary of the diverse findings from the evaluations of the Comprehensive Employment and Training Act (CETA) that were obtained when a number of analysts used diverse nonexperimental methods to evaluate the program.

[18] I was involved in the National JTPA study as a subcontractor on the component that investigated the possibility of using nonexperimental approaches to determine the impact of the program rather than experimental approaches.

[19] The final report was published as Orr et al. (1996).

ticipating sample of 16 SDAs (by design), but the report's overall conclusion is more optimistic: "The most basic conclusion . . . is that the study sites and the 17,026 members of the 18-month study sample resemble SDAs and their participants nationally and also include much of their diversity" (Bloom et al. 1993, p. 73).

Although the external validity of the National JTPA Study has been subject to a great deal of debate among analysts, there is no way to resolve the issue. Obviously it is best to avoid sites refusing to participate, but that may be easier said than done. Potential strategies to improve participation include larger incentive payments, exemption from performance standards sanctions for the period of participation,[20] making participation in evaluations mandatory in authorizing legislation, and decreasing the proportion of the applicants assigned to the control group.

7.2 Random assignment by service strategy recommended

Experimental methods can only be used to evaluate hypotheses where random assignment was used to assign the specific treatment received. In JTPA, the evaluators determined that prior to the experiment adults in the 16 sites were assigned to one of three broad categories – (1) occupational classroom training, (2) job search assistance (JSA) or on-the-job training (OJT), and (3) other services. Although OJT is generally the most expensive service strategy, because the program pays up to one-half of the participant's wages for up to six months, and JSA is the least expensive because it is generally of short duration and is often provided in a group setting, it was observed that the individuals deemed appropriate for OJT were virtually job ready as were those recommended for JSA; in addition, because OJT slots are difficult to obtain, candidates for OJT are often given JSA while waiting for an OJT slot to become available. The "other" category included candidates recommended for services such as basic skills (education), work experience, and other miscellaneous services but not occupational classroom training or OJT.

The strategy used in the National JTPA Study was to perform random assignment after a prospective participant was given a preliminary assessment and a service strategy recommended for the person; individuals that the program elected not to serve were excluded from the experiment. Two-thirds of the participants recommended for services were in the treatment group, and one-third was excluded from the JTPA program for a period of 18 months. During the embargo period, control group members were permitted to enroll in any workforce activities other than JTPA that they wished.

There are several concerns with the random assignment procedures used in the National JTPA Study. None of these concerns threatens the internal validity of the impacts estimated, but they show how difficult it is to test the most interesting hypotheses when trying to graft a random assignment experimental design to an existing program.

- *By presenting findings primarily per assignee rather than per participant, the findings may be misinterpreted*. This issue relates more to presentation than analysis. A reader of the full report can find detailed information about what the findings mean, but the executive summary stresses impact estimates per assignee, so casual readers may not learn the impact per person who enrolls in the program.[21] There are often large differences between the impact per assignee and impact per enrollee because for some analyses the percentage of assignees that actually enrolled in the program is much less than 100%. For adult women for example, less than half (48.6%) of the women assigned to classroom training actually received classroom training; for men, the figure was even lower (40.1%). Assignees who did not receive the recommended treatment strategy sometimes received other strategies, and the report notes that impacts per enrollee "were about 60 percent to 70 percent larger than impacts per assignee, depending on the target group" (Bloom et al. 1993, p. xxxv). Policy makers generally think about what returns they are getting on people who enroll in the program, as little, if any, money is spent on no-shows. Thus, policy makers want to know the impact per enrollee, and they might assume that impact estimates are impact per enrollee rather than impact per assignee.[22,23]
- *Failure to differentiate between the in-program period and the post-program period can be misleading, particularly for short-term findings.* The impact findings are generally presented on a quarterly basis, measured in calendar quarters after random assignment, or for the entire six-quarter follow-up period. For strategies that typically

[20] Although exempting participating sites from performance standards sanctions may increase participation, it also reduces external validity because the participating sites no longer face the same performance incentives.

[21] Some tables in the executive summary (e.g., Exhibit S.2 and Exhibit S.6) only provide the impact per assignee, and significance levels are only provided for estimates of impact per assignee.

[22] A U.S. Department of Labor senior official complained to me that one contractor refused to provide her with impacts per enrollee because they were based on nonexperimental methods and could not, therefore, be believed. She opined that the evaluation had little value for policy decisions if the evaluation could not provide the most important information she needed.

[23] Although I argue that estimates on the eligible population, sometimes referred to as "intent to treat" (ITT) estimates are prone to misinterpretation, estimating participation rates and the determinants of participation can be valuable for policy officials to learn the extent to which eligible individuals are participating and what groups appear to be underserved. See Heckman and Smith (2004).

last for more than one quarter, the reader can easily misinterpret the impact findings when the in-program and post-program impacts are not presented separately.[24,25]

- *The strategy does not allow head-to-head testing of alternative strategies.* Because random assignment is performed after a treatment strategy is recommended, the only experimental estimates that can be obtained are for a particular treatment versus control status. Thus, if, say, OJT has a higher experimental impact than classroom training, the experiment tells us nothing about what the impact of OJT would be for those assigned to classroom training. The only way to experimentally test this would be to randomly assign participants to treatment strategies. In the case of the JTPA, this would mean sometimes assigning people to a strategy that the SDA staff believed was inappropriate.
- *The strategy does not provide the impact of receiving a particular type of treatment – it only provides the impact of being assigned to a particular treatment stream*. If all JTPA participants received the activities they were initially assigned to, this point would not be important, but this was not the case. Among the adult women and men who received services, slightly over one-half of those assigned to occupational classroom training received this service, 58 and 56%, respectively.[26] Of those who did not receive occupational classroom training, about one-half did not enroll, and the remainder received other services. The figures are similar for the OJT-JSA group except that over 40% never enrolled. The "other services" group received a variety of services with no single type of service dominating. There is, of course, no way to analyze actual services received using experimental methods, but the fact that a relatively large proportion of individuals received services other than those recommended makes interpretation of the findings difficult.
- *The OJT-JSA strategy assignee group includes those receiving the most expensive services and those receiving the least expensive services, so the impact estimates are not particularly useful.* The proportions receiving JSA and OJT are roughly equal, but by estimating the impact for these two service strategies combined, policy and program officials cannot determine whether one of the two strategies or both are providing the benefits. It is impossible to disentangle the effects of these two very different strategies using experimental methods. In a future experiment this problem could be avoided by establishing narrower service strategies, e.g., making OJT and JSA separate strategies.
- *Control group members were barred from receiving JTPA services, but many received comparable services from other sources, making the results difficult to interpret.* The National JTPA Study states that impact estimates of the JTPA program are relative to whatever non-JTPA services the control group received. Because both the treatment group and the control group were motivated to receive workforce services, it is perhaps not surprising that for many of the analyses the control group received substantial services. For example, for the men recommended to receive occupational classroom training, 40.1% of the treatment group received such training, but so did 24.2% of the control group. For women, 48.6% of the treatment group received occupational classroom training and 28.7% of the control group received such services. Thus, to some extent, the estimated impacts do not provide the impact of training versus no training, but of one type of training relative to another.

The point is not that the National JTPA Study was seriously flawed; on the contrary, Hollister (2008) is correct to identify this study as one of the better social experiments conducted in recent years. Rather, the two key lessons to be drawn from the study are as follows:

- It is important to present impact estimates so that they answer the questions of primary interest to policy makers. This means clearly separating in-program and post-program impact findings and giving impacts per enrollee more prominence than impacts per assignee.[27]
- Some of the most important evaluation questions may be answered only through nonexperimental methods rather than experimental methods. Although experimental estimates are preferred when they are feasible, nonexperimental methods should be used when they are not. The U.S. Department of Labor has sometimes shied away from having researchers use nonexperimental methods in conjunction with experiments. When experimental methods cannot answer all the questions of interest, nonexperimental methods should be tried, with care taken to describe all assumptions made and for sensitivity analyses to be conducted.

[24] It is, of course, important to capture the impacts for the in-program period so that a cost-benefit analysis can be conducted.

[25] For example, Stanley et al. (1998) summarize the impact findings from the National JTPA Study by presenting the earnings impacts in the second year after random assignment, which is virtually all a post-program period.

[26] See Exhibit 3.18 of Bloom et al. (1993).

[27] This is not a simple matter when program length varies significantly, as it did in the JTPA program. If the participants are followed long enough, however, part of the follow-up period should be virtually should all be after program exit.

8 Conclusions

This paper has addressed the strengths and weaknesses of social experiments. There is no doubt that experiments offer some advantages over nonexperimental evaluation approaches. Major advantages include the fact that experiments avoid the need to make strong assumptions about potential explanatory variables that are unavailable for analysis and the fact that experimental findings are much easier to explain to skeptical policy makers. Although there is growing literature testing how well nonexperimental methods replicate experimental impact estimates, there is no consensus on the extent to which positive findings can be generalized.

But experiments are not without problems. The key point of this paper is that any impact evaluation, experimental or nonexperimental in nature, can have serious limitations. First, there are some questions that experiments generally cannot answer. For example, experiments frequently have "no-shows" who do not participate in the intervention after they were randomly assigned to the treatment group, and crossovers who are members of the control group who somehow take the treatment intervention or something other than what was intended for the control group. Experiments are often bad at capturing entry effects and general equilibrium effects.

In addition, in implementing experimental designs, things can go wrong. Examples include problems with participants understanding the intervention and difficulties in testing the hypotheses of most interest. These points were illustrated by showing how the National JTPA Study, which included random assignment to treatment status and is considered by many as an example of a well conducted experiment, failed to answer many of the questions of interest to policy makers.

Thus, social experiments have many advantages, and one should always give careful thought to using random assignment to evaluate interventions of interest. It should be recognized, however, that simply conducting an experiment is not sufficient to assure that important policy questions are answered correctly. In short, an experiment is not a substitute for thinking.

Executive summary

It is widely agreed that randomized controlled trials – social experiments – are the gold standard for evaluating social programs. There are, however, important issues that cannot be tested using experiments, and often things go wrong when conducting experiments. This paper explores these issues and offers suggestions on dealing with commonly encountered problems. There are several reasons why experiments are preferable to nonexperimental evaluations. Because it is impossible to observe the same person in two states of the world at the same time, we must rely on some alternative approach to estimate what would have happened to participants had they not been in the program.

Nonexperimental evaluation approaches seek to provide unbiased and consistent impact estimates, either by developing comparison groups that are as similar as possible to the treatment group (propensity score matching) or by using approaches to control for observed and unobserved variables (e.g., fixed effects models, instrumental variables, ordinary least squares regression analysis, and regression discontinuity designs). Unfortunately, all the nonexperimental approaches require strong assumptions to assure that unbiased estimates are obtained, and these assumptions are not always testable. Overall, the evidence indicates that nonexperimental approaches generally do not do a good job of replicating experimental estimates and that the most common problem is the lack of suitable data to control for key differences between the treatment group and comparison group. The most promising nonexperimental approach appears to be the regression discontinuity design, but this approach requires a much larger sample size to obtain the same amount of precision as an experiment.

Although a well designed experiment can eliminate internal validity problems, there are often issues regarding external validity. External validity for the eligible population is threatened if either the participating sites or individuals volunteer for the program rather than are randomly assigned. Experiments typically randomly assign people to the treatment or control group *after* they have applied for or enrolled in the program. Thus, experiments typically do not pick up any effects the intervention might have that encourage or discourage participation. Another issue is the finite time horizon that typically accompanies an experiment; if the experiment is offered on a temporary basis and potential participants are aware of the finite period of the experiment, their behavior may be different than if the program were permanent. Experiments frequently have no-shows and crossovers, and these phenomena can only be addressed by resorting to nonexperimental methods. Finally, experiments generally cannot capture scale or general equilibrium effects.

Several things can go wrong in implementing an experiment. First, the intervention might change while the experiment is implemented. A common occurrence is that the intervention itself changes, either because the original design was not working or circumstances change. The intervention should be carefully monitored to observe this and the evaluation modified if it occurs. Another potential problem is that participants may not understand the intervention; to guard against this, knowledge should be tested and instruction provided if it is a problem.

Many of the problems described here occurred in the random assignment evaluation of the Job Training Partnership

Act evaluation in the United States. Although the intent was to include a random sample of local programs, most local programs refused to participate, resulting in questions of external validity. Random assignment in the study occurred after an appropriate service strategy was selected. This assured that each strategy could be compared to exclusion from the program, but the alternative strategies could not be compared with each other. Crossover and no-show rates were high in the study, and it is likely many policy officials did not interpret the impact findings correctly. For example, 40% of the men recommended for classroom training received that treatment, as did 24% of the men in the control group. Thus, the difference in outcomes for the treatment and control groups is very different from the impact of receiving training versus not receiving training. Another feature that makes interpretation difficult is that one service strategy included those who received the most expensive strategy, on-the-job training, and the least expensive strategy, job search assistance; this makes it impossible to differentiate the impacts of these disparate strategies. Finally, the interim report made it difficult for the reader to separate impacts from the post-program period from those from the in-program period and much more attention was paid to the impact for the entire treatment group than the nonexperimentally estimated impact on the treated. It is likely that policy makers failed to understand the subtle but important differences here.

There is no doubt that experiments offer many advantages over nonexperimental evaluations. However, many problems can and do arise, and an experiment is not a substitute for thinking.

Kurzfassung

Es herrscht weitestgehend Konsens darüber, dass randomisierte kontrollierte Studien – Sozialexperimente – der „Goldstandard“ für die Bewertung sozialer Programme sind. Es gibt jedoch viele wichtige Aspekte, die sich nicht durch solche Studien bewerten lassen, und bei der Durchführung dieser Studien kann oft etwas schiefgehen. Die vorliegende Arbeit untersucht diese Themen und bietet Lösungsvorschläge für häufig entstehende Probleme. Es gibt viele Gründe, warum Experimente gegenüber nichtexperimentellen Bewertungen bevorzugt werden. Da es nicht möglich ist, die gleiche Person in zwei verschiedenen Zuständen gleichzeitig zu beobachten, müssen wir auf eine alternative Vorgehensweise zurückgreifen, um einzuschätzen, was mit den Probanden geschehen wäre, hätten sie am Maßnahmenprogramm nicht teilgenommen.

Nichtexperimentelle Bewertungsansätze versuchen unvoreingenommene, konsistente Aussagen über Auswirkungen zu treffen, indem sie entweder Vergleichsgruppen entwickeln, die der Behandlungsgruppe so ähnlich wie möglich sind („propensity score matching“), oder indem sie Ansätze verwenden, die beobachtete und nichtbeobachtete Variablen kontrollieren (z. B. Fixed-effects-Modelle, Instrumentalvariablen, „Ordinary Least Squares Regression Analysis“ und „Regression Discontinuity Designs“). Leider benötigen sämtliche nichtexperimentellen Ansätze starke Annahmen, um zu gewährleisten, dass unvoreingenommene Einschätzungen erfolgen. Es ist nicht immer möglich, solche Annahmen zu prüfen. Im Allgemeinen deuten alle Anzeichen darauf hin, dass nichtexperimentelle Ansätze nur schlecht experimentelle Einschätzungen reproduzieren können. Das häufigste Problem ist dabei der Mangel an geeigneten Daten, um die Kernunterschiede zwischen der Treatmentgruppe und der Vergleichsgruppe zu kontrollieren. Der vielversprechendste nichtexperimentelle Ansatz scheint das „Regression Discontinuity Design“ zu sein, wobei diese Methode eine wesentlich größere Versuchsgruppe benötigt, um die gleiche Präzision wie ein Experiment zu erreichen.

Obwohl ein gut geplantes Experiment Probleme der internen Validität ausschließen kann, bleiben oft Fragen der externen Validität. Die externe Validität hinsichtlich der Gesamtbevölkerung wird gefährdet, wenn entweder die teilnehmenden Standorte oder die Personen sich für das Programm freiwillig melden, anstatt zufällig ausgewählt zu werden. Normalerweise werden in Experimenten Personen zufällig der Treatmentgruppe oder der Kontrollgruppe zugeordnet *nachdem* sie sich für das Programm angemeldet haben. Auf dieser Weise bilden Experimente in der Regel Faktoren nicht ab, die Personen zur Teilnahme ermutigen oder von der Teilnahme abschrecken können. Ein weiterer Aspekt ist der begrenzte Zeithorizont, den ein Experiment normalerweise mit sich bringt. Läuft das Experiment nur für eine begrenzte Zeit und sind sich die potenziellen Teilnehmer dessen bewusst, kann ihr Verhalten anders sein, als wenn das Experiment zeitlich unbegrenzt wäre. Bei Experimenten muss man oft mit No-Shows und Cross-Overs rechnen, und nur nichtexperimentelle Methoden sind dafür geeignet, solche Phänomene zu berücksichtigen. Zuletzt können Experimente in der Regel Skaleneffekte und allgemeine Gleichgewichtseffekte nicht erfassen.

Bei der Durchführung von Experimenten kann einiges schief gehen. Erstens kann sich während der Durchführung die Intervention ändern. Dies passiert häufig, entweder weil das ursprüngliche Design sich als ungeeignet erwiesen hat oder weil sich die Bedingungen geändert haben. Die Intervention ist aus diesem Grund sorgfältig zu beobachten und die Bewertung gegebenenfalls entsprechend anzupassen. Ein weiteres potenzielles Problem ist die Möglichkeit, dass die Teilnehmer die Intervention nicht verstehen. Um hier vorzubeugen, sollten das Verständnis der Teilnehmer hinsichtlich der Intervention geprüft und ggf. Schulungen bereitgestellt werden.

Viele der hier beschriebenen Probleme sind bei randomisierten Bewertung des Job Training Partnership Act in

den USA aufgetreten. Obwohl eine Zufallsauswahl von lokalen Programmen teilnehmen sollte, weigerten sich die meisten dieser Programme. Diese Weigerung wirft Fragen der externen Validität der Studie auf. Die Randomisierung für die Studie erfolgte, nachdem eine passende Maßnahmenstrategie für die verschiedenen Teilnehmer ausgewählt worden war. Diese Vorgehensweise stellte sicher, dass jede Strategie mit der Situation bei Nichtteilnahme am Programm verglichen werden konnte, jedoch konnten die alternativen Strategien dadurch nicht miteinander verglichen werden. Die Cross-Overs und No-Show-Raten für die Studie waren hoch, und es ist wahrscheinlich, dass viele Beamte die Ergebnisse falsch interpretierten. Zum Beispiel bekamen nur 40% der Männer, für die eine Schulung empfohlen wurde, dieses Treatment, aber auch 24% der Männer in der Kontrollgruppe. Die unterschiedlichen Ergebnisse der Treatment- und Kontrollgruppen sind also nicht auf die Tatsachte zurückzuführen, dass eine Gruppe Schulungen bekommen hat und die andere nicht. Eine weitere Besonderheit, die die Interpretation schwierig macht, ist, dass eine Maßnahmenstrategie sowohl die teuersten Maßnahmen (die Ausbildung am Arbeitsplatz) als auch die billigsten Maßnahmen (die Hilfe bei der Jobsuche) enthielt. Dadurch ist es nicht möglich, zwischen den Auswirkungen dieser disparaten Maßnahmen zu unterscheiden. Schließlich machte es der Zwischenbericht dem Leser schwer, die Auswirkungen, die in der Zeit nach dem Programm beobachtet wurden, von denen während der Programmzeit zu trennen, und die Auswirkungen für die gesamte Treatmentgruppe bekamen viel mehr Aufmerksamkeit als die nichtexperimentell geschätzten Auswirkungen auf die Maßnahmenteilnehmer. Höchstwahrscheinlich sind den Entscheidungsträgern subtile, aber wichtige Unterschiede hier entgangen.

Es gibt keinen Zweifel, dass Experimente zahlreiche Vorteile gegenüber nichtexperimentellen Bewertungen haben. Es können dabei jedoch viele Probleme auftreten, und ein Experiment kann das Nachdenken nicht ersetzen.

Acknowledgements I am grateful to Laura Langbein, David Salkever, Peter Schochet, Gesine Stephan, and participants in workshops at George Washington University and the University of Maryland at Baltimore County for comments. I am particularly indebted to Jeffrey Smith for his thoughtful detailed comments and suggestions. Responsibility for remaining errors is mine.

References

Angrist, J.D., Krueger, A.B.: Does compulsory attendance affect schooling and earnings? Q. J. Econ. **106**(4), 979–1014 (1991)

Angrist, J.D., Krueger, A.B.: Instrumental variables and the search for identification: from supply and demand to natural experiments. J. Econ. Perspect. **15**(4), 9–85 (2001)

Barnow, B.S.: The impacts of CETA programs on earnings: a review of the literature. J. Hum. Resour. **22**(2), 157–193 (1987)

Barnow, B.S.: The ethics of federal social program evaluation: a response to Jan Blustein. J. Policy Anal. Manag. **24**(4), 846–848 (2005)

Barnow, B.S., Cain, G.G., Goldberger, A.S.: Issues in the analysis of selection bias. In: Stromsdorfer, E.W., Farkas, G. (eds.) Evaluation Studies Review Annual, vol. 5. Sage Publications, Beverly Hills (1980)

Bloom, H.S.: Accounting for no-shows in experimental evaluation designs. Evaluation Rev. **8**(2), 225–246 (1984)

Bloom, H.S., Orr, L.L., Cave, G., Bell, S.H., Doolittle, F.: The National JTPA Study: Title II-A Impacts on Earnings and Employment at 18 Months. Abt Associates, Bethesda, MD (1993)

Bloom, H.S., Michalopoulos, C., Hill, C.J., Lei, Y.: Can Nonexperimental Comparison Group Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare to Work Programs? MDRC, New York (2002)

Blustein, J.: Toward a more public discussion of the ethics of federal social program evaluation. J. Policy Anal. Manag. **24**(4), 824–846 (2005a)

Blustein, J.: Response. J. Policy Anal. Manag. **24**(4), 851–852 (2005b)

Burtless, G.: The case for randomized field trials in economic and policy research. J. Econ. Perspect. **9**(2), 63–84 (1995)

Card, D., Robins, P.K.: How important are "entry effects" in financial incentive programs for welfare recipients? J. Econometrics **125**(1), 113–139 (2005)

Connell, J.P., Kubisch, A.C.: Applying a theory of change approach to the evaluation of comprehensive community initiatives: progress, prospects, and problems. In: Fulbright-Anderson, K., Kubisch, A.C., Connell, J.P. (eds.) New Approaches to Evaluating Community Initiatives, vol. 2, Theory, Measurement, and Analysis. The Aspen Institute, Washington, DC (1998)

Cook, T.D., Shadish, W.R., Wong, V.C.: Three conditions under which experiments and observational studies produce comparable causal estimates: new findings from within-study comparisons. J. Policy Anal. Manag. **27**(4), 724–750 (2008)

Dehejia, R.H.: Practical propensity score matching: a reply to Smith and Todd. J. Econometrics **125**(1), 355–364 (2005)

Dehejia, R.H., Wahba, S.: Propensity score matching methods for nonexperimental causal studies. Rev. Econ. Statistics **84**(1), 151–161 (2002)

Goldberger, A.S.: Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations. Institute for Research on Poverty, Discussion Paper 123–72, University of Wisconsin, Madison, WI (1972)

Greenberg, D.H., Shroder, M.: The Digest of Social Experiments, 3rd edn. The Urban Institute Press, Washington DC (2004)

Hahn J., Todd, P.E., Van der Klaauw, W.: Identification and estimation of treatment effects with a regression discontinuity design. Econometrica **69**(1), 201–209 (2001)

Heckman, J.J.: Economic causality. Int. Stat. Rev. **76**(1), 1–27 (2008)

Heckman, J.J., Smith, J.A.: Assessing the case for social experiments. J. Econ. Perspect. **9**(2), 85–110 (1995)

Heckman, J.J., Smith, J.A.: The determinants of participation in a social program: evidence from a prototypical job training program. J. Labor Econ. **22**(2), 243–298 (2004)

Heckman, J.J., Ichimura, H., Todd, P.E.: Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. Rev. Econ. Stud. **64**(4), 605–654 (1997)

Heckman, J.J., Hohmann, N., Smith, J., Khoo, M.: Substitution and dropout bias in social experiments: a study of an influential social experiment. Q. J. Econ. **115**(2), 651–694 (2000)

Hollister, R.G. jr.: The role of random assignment in social policy research: opening statement. J. Policy Anal. Manag. **27**(2), 402–409 (2008)

Lise, J., Seitz, S., Smith, J.: Equilibrium Policy Experiments and the Evaluation of Social Programs. Unpublished manuscript (2005)

Moffitt, R.: Evaluation methods for program entry effects. In: Manski, C., Garfinkel, I. (eds.) Evaluating Welfare and Training Programs. Harvard University Press, Cambridge, MA (1992)

Orr, L.L.: Social Experiments: Evaluating Public Programs with Experimental Methods. Sage Publications, Thousand Oaks, CA (1999)

Orr, L.L., Bloom, H.S., Bell, S.H., Doolittle, F., Lin, W.: Does Training for the Disadvantaged Work? Evidence from the National JTPA Study. The Urban Institute Press, Washington, DC (1996)

Rolston, H.: To learn or not to learn. J. Policy Anal. Manag. **24**(4), 848–849 (2005)

Schochet, P.Z.: National Job Corps Study: Methodological Appendixes on the Impact Analysis. Mathematical Policy Research, Princeton, NJ (2001)

Schochet, P.Z.: Comments on Dr. Blustein's paper, toward a more public discussion of the ethics of federal social program evaluation. J. Policy Anal. Manag. **24**(4), 849–850 (2005)

Schochet, P.Z.: Statistical power for regression discontinuity designs in education evaluations. J. Educ. Behav. Stat. **34**(2), 238–266 (2009)

Shadish, W.R., Clark, M.H., Steiner, P.M.: Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. J. Am. Stat. Assoc. **103**(484), 1334–1343 (2008)

Smith, J.A., Todd, P.E.: Does matching overcome LaLonde's critique of nonexperimental estimators? J. Econometrics **125**(1), 305–353 (2005a)

Smith, J.A., Todd, P.E.: Rejoinder. J. Econometrics **125**(1), 305–353 (2005b)

Stanley, M., Katz, L., Krueger, A.: Developing Skills: What We Know about the Impacts of American Employment and Training Programs on Employment, Earnings, and Educational Outcomes. Cambridge, MA, unpublished manuscript (1998)

Wilde, E.T., Hollister, R.: How close is close enough? Evaluating propensity score matching using data from a class size reduction experiment. J. Policy Anal. Manag. **26**(3), 455–477 (2007)

Wilson, L.A., Stoker, R.P., McGrath, D.: Welfare bureaus as moral tutors: what do clients learn from paternalistic welfare reforms? Soc. Sci. Quart. **80**(3), 473–486 (1999)

Burt S. Barnow is the Amsterdam Professor of Public Service at the Trachtenberg School of Public policy and Public Administration at George Washington University. Dr. Barnow has over 30 years of experience as an economist and manager of research projects in the fields of workforce investment, program evaluation, performance analysis, labor economics, welfare, poverty, child support, and responsible fatherhood programs. Prior to coming to George Washington University, Dr. Barnow was Associate Director for Research at Johns Hopkins University's Institute for Policy Studies, where he worked for 18 years. Prior to that, he worked for 8 years at the Lewin Group and nearly 9 years at the U.S. Department of Labor, including 4 years as Director of the Office of Research and Evaluation in the Employment and Training Administration. Prior to those positions, Dr. Barnow was an assistant professor of economics at the University of Pittsburgh. He has a B.S. degree in economics from the Massachusetts Institute of Technology and M.S. and Ph.D. degrees in economics from the University of Wisconsin at Madison.