

13462

July 1994

EVALUATING SOCIAL POLICIES: PRINCIPLES AND U.S. EXPERIENCE

Jean Baldwin Grossman

FILE COPY

Invariably, studies, proposals, and plans for social programs contain a strong recommendation for evaluation and monitoring. Reliable information about what works and why is clearly vital for improving existing programs or designing future ones. Making such assessments requires effective methods of evaluation. Policymakers who use these evaluations need to know about the methods—the pitfalls to watch for and the relative advantages and disadvantages of different techniques in different situations. This article describes these evaluation methods and the experience accumulated in the United States in applying them in practice.

Does a particular social program have the effect it is intended to have? If not, why not? Policymakers need answers to these questions if they are to make the most effective use of limited resources to advance social goals. In developing countries in particular, governments and development institutions cannot afford to waste scarce resources on programs that do not achieve their goals. Ineffective programs should be modified to make them work better or be canceled altogether. Thus, evaluating or monitoring performance is often strongly recommended and may even be a requirement of a program's funding.

But policymakers often know very little about *how*—and equally important, *how not*—to assess a program's effectiveness, and therefore cannot influence or adequately judge the quality of the evaluation methods proposed. Every newly initiated program is designed and fully expected to work. Most program designers, funders, and operators believe their programs do help. They can always point to particular individuals or communities whose conditions improved. Anecdotal evidence is often very persuasive, especially if one is inclined

to believe a particular program is working—as a program’s creator would be. But did the program cause the improvement? Would the situation of these individuals or communities have improved anyway, with help from some other source? Did the size or extent of the improvement justify the program’s cost?

An evaluation technique that can accurately address these issues arms policymakers with the information they need to determine whether a program is worth the cost. For example, evaluations of a nutrition program for pregnant women, infants, and children in the United States established that the program was indeed saving the country substantial amounts of money: “For every dollar spent on [the program], the associated saving in [government-funded medical costs] during the first sixty days [postpartum] ranged between \$1.77 to \$3.13” (Devaney, Bilheimer, and Shore 1992, Executive Summary). Similarly, an evaluation of a residential education and training program for high school dropouts in the United States found that, even though the program was relatively expensive, society more than recouped the costs through increased work effort by the participants, increased tax revenues, lower crime rates, and lower social service costs (Mallar and others 1982). Findings such as these from rigorous evaluations provide a reliable foundation on which to base decisions on whether to continue, modify, or terminate a program. These decisions can save countries or communities money that otherwise would be wasted on a mistargeted program or lost in closing down a program whose cost-effectiveness was not immediately apparent.

A rigorous evaluation can also guide the modification needed to improve program delivery. For example, during a particularly deep and long recession, the U.S. government often provides unemployed workers with extra unemployment compensation. An evaluation of such an emergency compensation program in the late 1970s (Brewster and others 1978) found that workers did not actually receive much of the money until after the economy began to recover. The rules for distributing the money were accordingly changed when a compensation scheme was next needed during the recession of the early 1980s. An evaluation of the modified program indicated that the new rules allowed policymakers to deliver the aid more precisely when it was needed most.

How does one set up an evaluation that will generate reliable and useful information? The answer can be informed by what has been learned elsewhere. For policymakers and evaluators who must weigh the advantages and disadvantages of alternative techniques, U.S. experimentation in this area can provide valuable guidance. Researchers in the United States have expanded and refined the theory of evaluation and have field-tested several different methods. This article draws on this accumulated experience to lay out some of the strategies that have been developed, describing why some have fallen out of favor and why the value of others is being questioned. The analysis goes on to consider what conditions are conducive to each of the principal methods currently in use.¹

Evaluation Techniques

A program's effect can be measured accurately only if one knows what would have happened without it. Because one obviously cannot observe the outcomes for the participants themselves had they not enrolled in the program, a proxy group of nonparticipants must be identified. Determining this hypothetical no-treatment (or counterfactual) state is the crux of designing an evaluation because, under any strategy, a program's effects are ascertained by comparing the behavior of the treatment or participant group with the behavior of the selected counterfactual group.

Indeed, determining the no-treatment state is so central to an evaluation that designs are categorized according to the way in which the counterfactual group is selected—*nonrandom assignment* (classified as a quasi-experimental design) and *random assignment* (classified as an experimental design). In the quasi-experimental category the two principal types of design are *reflexive techniques*, in which the postprogram behavior of participants is compared with their preprogram behavior, and *matched comparisons*, in which the postprogram behavior of the participants is compared with the behavior of a group of individuals who were similar to the participants before they enrolled in the program. In this article, as in much other research, the word “controls” is reserved to mean members of a randomized control group. All other counterfactual groups are called “comparison groups.”

Quasi-Experimental Designs

This section discusses the design of the two major types of quasi-experimental method—reflexive and matched comparison—with examples of evaluations that have used each method. For each, the way the counterfactual group is constructed is pivotal.

REFLEXIVE COMPARISON. In a reflexive comparison the participants serve as both the treatment and the comparison group. The counterfactual state is surmised using the preprogram behavior of the participants to infer what would have happened to them had they not joined the program. The strength of this methodology is that the socioeconomic and demographic characteristics of the group, previous experience, and the individuals' predisposition and innate abilities are the same both before and after the program. Consequently, observed changes in behavior from the pre- to postprogram period cannot be attributed to differences in these factors. Careful mathematical modeling is, however, required to ensure that changes that would have occurred naturally are not attributed to the program. Generally, reflexive comparison group studies are time-series or panel studies that collect a large amount of data for several years both before and after the program to enable the researchers to understand how factors other than the program influence the outcome.

A reflexive comparison was used to evaluate the effect of a water conservation campaign in the United States. In 1972 a particular county declared a moratorium on new water hookups until alternative sources for water could be assessed. This moratorium lasted three years. Using monthly data from 1966 to 1976, researchers were able to estimate statistically that the moratorium reduced water consumption by 15 percent (Maki, Hoffman, and Berk 1978). The finding has been used to justify the passage of similar regulations during other drought periods and in other counties.

Because this technique requires a relatively long observation period before and after the program and is more dependent on statistical analysis and assumptions than matched comparison or random assignment, it is the least used of the three methods for evaluating social policy. The technique is more common in research (such as psychological studies) where the key factors affecting the outcomes—such as an individual's self-esteem or resilience to adversity—are very difficult to measure accurately and large research samples are not feasible.

MATCHED COMPARISON. In designing matched comparison groups, researchers identify a group of individuals whom the researchers judge to be comparable to the participant group in important dimensions but who do not receive program services. The researchers should match the two groups on factors that are known or believed to affect the key outcomes significantly. Such knowledge comes through previous experience or a theoretical understanding of the processes expected to underlie the intervention. The participant and comparison groups do not have to be similar with respect to characteristics that do not affect the outcomes of interest. For example, the researchers should draw on knowledge about factors that affect particular crops when selecting agricultural districts for comparison in a study of a program designed to affect agricultural productivity. The aggregate behavior of the comparison group is then assumed to indicate how the participants would have behaved had they not joined the program.

Two principal types of matched comparison groups are prospective studies, in which comparison group members are selected at the same time as participants are enrolling in the program, and retrospective studies, in which comparisons are selected at a single point after the participant group has been enrolled. An example of a prospective study is the evaluation of the California Conservation Corps (CCC), a training and environmental conservation program for out-of-school youth (Wolf, Liederman, and Voith 1987). In this study selected participants who enrolled in the program during a twelve-month period were inducted into the research sample. During the same period, similar individuals who went to the CCC's single largest referral agency were designated as comparison group members. Before the study began, researchers talked with referral staff to find out what kind of people were encouraged to enroll in the CCC. Then a brief survey was conducted at the referral agency and the CCC to determine how the people flowing through the two organizations

differed. On the basis of this survey, comparison group members were matched not only on age, race, and gender but on whether they had children, how receptive they were to moving away from home, how much they enjoyed working outdoors, and how much they enjoyed physical work. These were all important factors in corps members' decisions to join the program and might affect self-esteem, environmental awareness, and earnings—key outcomes of the study.

An example of a retrospective study is the evaluation of the Job Corps, an education and training program for out-of-school youth (Mallar and others 1982). In this evaluation, Job Corps participants were compared with a group of individuals who had been surveyed by the Census Bureau for the Current Population Survey (CPS). The comparison group individuals (hereafter termed comparisons) were matched to the participants with respect to age, race, gender, poverty status, and education. All these factors were important in predicting future income. To minimize the risk that comparisons had enrolled in a Job Corps (information not given in the CPS), they were selected only from areas that were not served by Job Corps centers.

The CCC and Job Corps studies are good examples of evaluations that have proved important in policy decisions. In particular, the Job Corps study found that the benefits to society outweighed the program's cost. Advocates used this finding to win political support for continued funding for the Job Corps, which had been slated for closure because of high costs. Similarly, advocates of the CCC have used the evaluation findings to secure funding during periods of budgetary cutbacks.

Matched comparison group methodology is the most common type of evaluation used in the United States, but since the 1970s the flaws in straightforward comparisons have been increasingly recognized. For example, individuals who do not participate in a program when given the opportunity may well differ considerably from those who do, in such attributes as their motivation or their ability to benefit from the program—factors that would positively affect desired outcomes of the program. Because such factors are not easy to measure, it is hard to ensure, even statistically, that they were similar for both groups. Thus, the relevant outcome (say, earnings) for the nonparticipants in this situation would be a poor proxy for the “would-have-been” outcome for the participants because the comparison group members are not comparable to the participants along important unobservable characteristics. This source of bias has come to be known as sample, or self-selection, bias.

Studies evaluating the Salk (polio) vaccine and the Manpower Development and Training Act (MDTA), the employment and training program undertaken in the United States in the 1960s, illustrate some of the pitfalls of these quasi-experimental (nonrandom) comparison group methods. The Salk vaccine was evaluated using two different techniques, one quasi-experimental (a comparison of vaccine recipients with a nonrandom, naturally occurring group of individuals who did not receive the vaccine) and the other experimental (a comparison

of randomized groups). The estimate of the vaccine's effectiveness derived from the data from the randomized experiments was 14 percent higher than the estimate derived from the nonrandomly generated data. The nonrandom MDTA evaluation (Westat, Inc. 1984) compared the earnings of participants with those of nonparticipants. The estimated effect showed that the program decreased the income of trainees, even when the researchers controlled for differences in the two groups. To investigate this unexpected result, Director (1974) examined the earnings of participants and comparison group individuals before the program and found that, even before the program, the participants had earned less than the comparison group members. Thus Westat's estimate of a negative impact was likely incorrect.

Reflexive evaluations that assume that a participant's situation before and after participating in the program would be similar may also not be valid. For example, if a string of bad luck with employment brought an individual to a training program, simply comparing the individual's situation before and after the program would overstate the effectiveness of a program because the participant's situation would have improved (on average) even in the absence of the program. Similarly, if the participant's situation were trending up or down, a before-and-after comparison would produce biased impact estimates. For example, a reflexive study of arthritis treatment found that the severity of the condition *worsened* over the life of the project. Randomized trials, however, showed that the treatment slowed the deterioration of the patient's condition (Deniston and Rosenstock 1972).

This threat of trending is especially important when the program serves young people. For example, the Summer Training and Education Program, a federal pilot program, was a summer remediation and work-experience program for educationally and economically disadvantaged 14- and 15-year-olds. During the first summer of the program, participants lost half a grade in reading ability. If a before-and-after comparison had been used to judge the program's effectiveness, it would have been concluded that the extra instruction had harmed the participants. Fortunately, however, the evaluation was a randomized experiment. The control group members lost a full grade in reading—indicating that the true effect of the program was to raise test scores more than half a grade (Sipe, Grossman, and Milliner 1987).

In the mid-1970s James Heckman salvaged the credibility of comparison group methodologies to some degree by developing a statistical solution enabling researchers to control theoretically for these unobservable differences (Heckman 1980). But many researchers felt uneasy about the statistical assumptions needed to correct for the biases and therefore searched for alternative solutions.

Experimental Designs

Using random assignment in social policy evaluation was one of the alternatives found. In a random assignment evaluation, individuals who are eligible

for a program are identified and then randomly assigned to one of two groups: a treatment group, which is offered the program, or a control group, which is not. This methodology ensures that, before the program, the two groups are statistically equivalent, on average, with respect to *all* characteristics, observed and unobserved. Thus, if the average behavior of the two groups differs after the intervention, the difference can be confidently and causally linked to the program. Impact estimates based on random assignment evaluations are far more defensible than estimates based on nonrandom designs so long as the behavior of all treatments and all controls is compared—including even those members of the treatment group who did not participate in the program, by chance or by choice—and so long as researchers check and control for potential sample bias arising from attrition.² With these safeguards, those who do not like the results (including the sponsors of the research and the evaluators themselves) cannot attack the evaluation by claiming that the behavior of the controls does not accurately reflect what would have happened to the treatment group members if they had not received the intervention.

One of the most recent and sophisticated examples of randomized social experiments is the evaluation (Bloom and others 1992) of the Job Training and Partnership Act (JTPA) program, a U.S. national training and employment program adopted in the 1980s. For this study, the first randomized evaluation of a national program, applicants were randomized after their training needs had been assessed so that a treatment and control comparison could be created for each type of service typically provided to JTPA participants—basic education, work experience, and on-the-job training (see Hotz 1992 for a detailed discussion of the evaluation design). An example from the 1970s is the Negative Income Tax (NIT) experiment, which was one of the first large randomized social experiments conducted in the United States. For this study, low-income families were randomly assigned different guaranteed minimum incomes and different tax rates on earnings. Despite the lack of a “no-guaranteed-income” group, comparisons among the different treatment groups enabled researchers to develop unbiased estimates of the program’s effect on participants’ incentive to work and their general well-being (see Kershaw and Fair 1976 for details on the design and implementation of the evaluation).

Both of these evaluations were multiyear, multisite, multimillion-dollar projects, and they had a significant effect on policy. The JTPA evaluation, which found few positive effects on participants’ income, helped convince many in the U.S. government of the need to revamp the program. Although, for political reasons, a negative income tax was never enacted, subsequent revision of the Food Stamp program incorporated the taxing scheme of the NIT.

In addition to these large experiments, many smaller randomized evaluations have been conducted. One of these assessed the effects of a culturally appropriate Mexican version of the children’s television program “Sesame Street.” The principal question in the Mexican study was whether the program increased knowledge among its viewers. The study found that children did

benefit, and “Plaza Sesamo” was consequently aired on the official government channel.

Experimental versus Quasi-Experimental Designs

Until the mid-1980s randomized and comparison group methodologies coexisted without too much controversy. Statisticians and econometricians continued to develop better ways of using comparison group data to estimate the effects of programs (Heckman and Robb 1985a, 1985b). However, the big drawback of quasi-experimental methods—the need for statistical adjustment—remains. Evaluators identify members of a comparison group on the basis of judgments informed by theory and experience. They then attempt to control for any remaining differences through statistical techniques. The validity of the selection process remains an empirical question that can be resolved only after the data are collected. Because impact estimates derived from randomized experiments need no statistical adjustment, they are inherently less controversial and easier to defend. During the 1970s and early 1980s most evaluators recommended the use of random assignment for evaluating major program initiatives but considered comparison group strategies acceptable for smaller evaluations, in view of advances that had been made in statistical modeling.

But then controversy over a comparison group evaluation of the Comprehensive Employment and Training Act (CETA) dealt a serious blow to researchers’ faith in statistical modeling. The U.S. government wanted to determine the effectiveness of the program, but politics dictated that no individual who desired services be denied them, which put a random assignment evaluation out of bounds. The government therefore funded a massive comparison group evaluation. Extensive data were collected on CETA participants nationally. The participants’ behaviors were then compared with the behaviors of individuals with similar characteristics identified in other national surveys. Comparison group members were matched to participants with respect to gender, race, age, education, poverty status, and income history. Sophisticated matching algorithms were used to select the most comparable individuals. The studies found surprisingly few positive effects of the program (Bassi and others 1984; Dickerson, Johnson, and West 1984; Westat, Inc. 1984). But the sensitivity of the estimates to the statistical techniques used was disturbing. The evaluation community began to view this study and its results with suspicion.

Fraker and Maynard (1984) wrote a seminal paper that showed how comparison group methodologies, such as the one used in the CETA study, could lead to false negative estimates of impacts. Using data from a major random assignment demonstration program conducted in the United States, the National Supported Work (NSW) demonstration, the authors constructed comparison groups using various matching strategies, including those employed in the CETA evaluations. The authors calculated what the impact estimates would have been had the outcomes for the NSW treatment group been compared with

the outcomes for the matched comparison groups, instead of with those for the randomly selected NSW control group. They found that the estimates based on comparison groups did not come close to the actual impacts estimated using the randomized control group.

Lalonde (1986) further damaged the case for comparison groups by showing that the various sophisticated econometric techniques developed to solve self-selection problems did not improve Fraker and Maynard's quasi-experimental estimates. Summarizing their findings, Lalonde and Maynard (1987) found that the statistical inference based on a comparison group methodology was incorrect 40 percent of the time. They concluded the following:

Nonexperimental procedures may not accurately estimate the true program impacts. In particular, there does not appear to be any formula (using nonexperimental methods) that researchers can confidently use to replicate the experimental results of the Supported Work Program. In addition, these studies suggest that recently developed methods for constructing comparison groups are no more likely (and arguably less likely) than the econometric procedures to replicate the experimental estimates of the impact of training. . . . These findings are further evidence that the current skepticism surrounding the results of nonexperimental evaluations is justified (Lalonde and Maynard 1987, p. 226).

These papers had a profound influence on evaluators' confidence in comparison group methodologies, ultimately undermining the CETA study so badly that most of the evaluation and policy communities ignored its results and discounted its conclusions. Thus the millions of dollars invested in the study were wasted because the findings lacked credibility.

It was this research that led a government-appointed technical advisory panel to recommend that a randomized design evaluation replace the comparison group evaluation originally planned for the JTPA, the program that succeeded CETA. The panel wrote that its recommendations

are strongly conditioned by the judgment that it will not be possible to solve the problem of selection bias within the context of a quasi-experimental design, at least not in [a] short enough time frame to meet Congress' need for valid information to guide policy. Even though many authors studying employment and training programs have recognized the selection problem, no such study using a quasi-experimental design can be said to have controlled adequately for selection bias. The panel does not intend to set forth a counsel of despair. Rather, it is concerned that the past evaluations of CETA have consumed, and the contemplated evaluations of JTPA will consume, millions of dollars and much valuable time. It would be extremely unfortunate if the analysis of [the originally planned JTPA evaluation design] would yield the same ambiguous conclusions as

has the analysis of the [quasi-experimental] data base for CETA (Job Training Longitudinal Survey Research Advisory Panel 1985, p. 21).

Nonetheless, the panel thought it "prudent to continue development work on the econometric front (p.22)." The "use of the quasi-experimental analysis design in the long run may be the only way one can regularly measure nationally representative program outcomes and be able to do this without having to deny program treatment to some (p.22)," the panel wrote. Thus the panel persuaded the government to fund an evaluation of comparison group methods under the leadership of James Heckman and Joseph Hotz.

Heckman and Hotz found the Lalonde-Fraker-Maynard argument flawed in that those researchers had used various econometric techniques without first testing to see if they were statistically appropriate. Using the same data, Heckman and Hotz conducted tests of these techniques and found that estimates leading to incorrect statistical inferences were based on techniques or specifications that could be rejected (Heckman, Hotz, and Dabos 1987). They concluded, "Given that a variety of robust nonexperimental methods have yet to be assessed and that better data on the selection process are being generated by a new JTPA evaluation study, which will shed important new light on the selection process, we are confident that reliable nonexperimental evaluation methods can and will be developed in the future for all subsidized employment and training programs (p.424)."

Friedlander and Robins (1992) applied the quasi-experimental specification testing strategy to experimental data from the work-welfare demonstrations conducted in the United States during the late 1980s to see if the tests could successfully weed out the techniques that had led to incorrect conclusions in another (non-NSW) setting. They found that even when specification tests were incorporated, the reliability of the quasi-experimental results did not improve much. Thirty-six percent of the estimates that had passed the tests still led to the wrong statistical inference, and 56 percent of those that had failed the tests and had therefore been discarded were correct. The authors found that the most egregiously incorrect impact estimates were indeed discarded but that many correct estimates were thrown out at the same time. In defense of the Heckman-Hotz position, however, it should be noted that Friedlander and Robins used a data base that did not explicitly collect information that allowed extensive specification tests and modeling of the selection process.

An important postscript to this discussion is that the effects found in the randomized JTPA evaluation echoed those found in the nonrandom CETA evaluations, whose credibility had been so thoroughly undermined. JTPA appears to have a modest positive effect on earnings for adults eighteen months after enrollment, no effect on out-of-school young women, and a negative effect on out-of-school young men. The weak components of JTPA are basically the same components that were ineffective in CETA. Had the original CETA evaluation been a random assignment evaluation, policymakers would have had to contend with its findings and change the program. Instead, the earlier results were

simply ignored and the same ineffective (or worse) services were repeated for an additional ten years.

Where does this discussion lead? I believe it leads to the conclusion that random assignment design is the evaluation strategy that yields the most reliable and defensible estimates. If the question at issue has profound implications for policy, it may pay to invest the time and resources necessary for random assignment (see Heckman 1992 for a somewhat technical discussion of the relevant benefits of randomized versus quasi-experimental designs). But random assignment is not appropriate, or even possible, in many situations, as the discussion that follows explains.

Choosing an Evaluation Strategy

Choosing an evaluation method depends on the kinds of questions being asked about a program, the number of participants the program will serve, the operational details that might make one or another type of evaluation unsuitable, and the constraints on the time and resources available for the exercise.

Key Questions

In selecting a technique suited to assessing a particular program, policymakers must know which questions they most want answered. One evaluation technique may be more suited to tackling certain types of questions than another. For instance, both quasi-experimental and experimental strategies can investigate the effects of a program on individuals—such as the effectiveness of a particular retraining program in speeding workers' transition to new employment. But if the principal questions concern the program's effects on the community, the providers of the service, or other aggregate organizations—for example, the likely effects on the local labor market of guaranteeing all young people a summer job—then experimental random assignment designs are not appropriate.

Often, the intended scale of the program dictates which of these sets of questions—individual or community—are the most important. For example, suppose that budget constraints allow only a fraction of the unemployed to receive retraining services. In such a case estimates of effects on individuals are likely to be of most interest, and any of the evaluation strategies could be appropriately applied. Or consider a farm subsidy program aimed at increasing the incomes of farmers. The principal questions again are about the program's effects on individuals rather than about the feasibility of large-scale implementation, and again any of the evaluation strategies would be suitable. But if the subsidy program is to serve every farmer who qualifies, and if issues such as the supply responses of the farmers are critical, then the evaluation options are more limited. To observe these aggregate or community-level responses to the

program, a trial demonstration of the program should simulate universal eligibility in certain sites (a technique known as "saturation"). Because random assignment strategies require that some eligible individuals do not receive the treatment, demonstrations that saturate a site must be evaluated with quasi-experimental methods. For the same reason, quasi-experimental strategies must be used to evaluate existing programs that serve everyone who meets the eligibility criteria, such as Social Security in the United States or a national health service.³ Existing full-coverage programs can be evaluated by examining different effects produced by different program intensities. Reflexive pre- and postprogram design can also be used to evaluate recently implemented programs. But in both cases statistical analysis must be conducted to control for intervening factors.

An example of a saturation demonstration is the Youth Incentive and Employment Project, in which young people were guaranteed a job if they remained in school (Farkas and others 1982). The key questions were whether enough jobs could be created to fulfill this promise and whether the program increased graduation rates. Matched comparison was used here, pairing cities with similar labor market and youth characteristics. The study found that enough jobs could be created and that, during the program, earnings among black youth increased to nearly equal earnings of white youth in the participating sites (a condition that did not exist before the program or in the comparison sites).

Another saturation demonstration, the Employment Opportunity Pilot Program (EOPP), was mounted by the U.S. government in the late 1970s to field-test an alternative welfare, employment, and training program. The EOPP provided intensive job search assistance to all "employable" welfare recipients and all "employable" members of low-income households (Brown and others 1984). Individuals who could not find a job were given subsidized employment or training opportunities. One of the critical questions was whether an entitlement program such as EOPP was operationally feasible. How many jobs would be needed? How would the program affect the general labor market for low-wage jobs? To answer these questions, the demonstration consisted of ten saturation sites, in each of which sufficient job and training positions were to be funded to satisfy the full demand. Ten comparison sites were chosen that were judged to have populations, industrial structures, and labor market conditions similar to the EOPP sites.

The original evaluation design also included an eleventh, random assignment, site. Information from this site would be used to gauge the validity of the estimated effects derived from the other sites. For budgetary and political reasons, this eleventh site was eliminated, but the attempt illustrates how multiple strategies can be used to strengthen the overall quality of the evaluation.

A potential problem with a matched site strategy is that the cities or sites that make up a matched pair may become dissimilar during the follow-up period. The longer the follow-up, the more likely the sites' situations are to

drift apart. For example, in selecting comparison sites in the EOPP demonstration, changes in the unemployment rates from 1975 to 1978 (a preprogram period) were examined to determine whether the pairs' economies were likely to respond similarly to the improvement expected in national economic conditions (Brown and others 1984). But the economy of one of the comparison sites depended heavily on the auto industry, which was hit much harder by the recession of the early 1980s than by the recession of the mid-1970s. As a result, the comparability of this comparison site with its designated demonstration site was compromised. The researchers were thus unable to control effectively for local labor market conditions; instead, results were estimated both with and without this problem site.

A slow phase-in of a program is conducive to mixed-mode evaluation. For example, consider a program that, because of financial or resource constraints, can be initiated in only a few sites and even then can serve only a small fraction of the individuals it is ultimately intended to serve. This situation opens the possibility of conducting a random assignment evaluation during the first few years of the program's operation to determine if it improves the lives of its participants. (For such an evaluation to be most useful, the limited early version of the program should be offered to the same set of individuals as those who will ultimately qualify for the program—not, however tempting, to those most in need, or to some other special group, because these would not be representative of the intended target population.) Once additional capacity is available and the program can be expanded to serve all intended recipients, more effects at the community level can be observed and analyzed through matched comparisons or some other quasi-experimental method.

Number of Participants

Some social programs are designed to affect individuals, others to affect communities or entire regions. Employment and training programs are geared primarily to individuals; health clinics affect communities; irrigation programs affect farms or regions; and most educational programs affect classes of young people. The number of potential program participants—millions of people, thousands of schools, hundreds of towns, a handful of regions—will affect the type of evaluation that is possible. When the number is small, matched comparison or reflexive techniques are more likely to produce a counterfactual group more similar to the participant group than random assignment. Random assignment strategies are best when the number of potential participants is fairly large.

The earlier discussion of matched comparison groups described how evaluators could carefully create a comparison group by selecting nonparticipants with identical sets of key characteristics, such as age and education (or rainfall and soil richness in an agricultural example). Random assignment obviously cannot exactly match individuals in this way; for example, the likelihood that

two randomly chosen individuals will be the same age is quite small. So, if key characteristics of a participant group and a comparison group can be matched exactly through matched comparison, why would one ever use random assignment? The reason is that although the characteristics explicitly matched by the researcher are the same, other, unobservable characteristics that might affect the outcome of the program are not.

The question, then, is how to create similar groups using random selection. Two randomly chosen individuals are unlikely to be the same age, but the average age of two randomly selected groups of a thousand people is likely to be quite close. In fact, the average of all characteristics (both observable and unobservable) of these two large groups—education level, IQ, healthiness, work history—is likely to be quite similar. If individuals are randomly assigned to groups, as the size of the groups increases, the average characteristics of the two groups converge. The two groups are unlikely ever to be identical, but statisticians can calculate how large the groups must be to be confident that any small observable difference is due to chance. Random assignment, thus, is a more reliable means of ensuring that, before the program begins, the control group is similar to the treatment group, on average, on *all* characteristics. Any differences between the two groups observed after the program can thus be attributed to the program.

How large the groups of participants and controls must be for a useful evaluation depends on how similar individuals are to one another naturally, at least with respect to the key outcome, and how much the program is expected to change this outcome (see Conlisk 1973 for a discussion of the underlying assumptions, specific formulas, and more general forms of the problem). If the outcome does not vary at all across the target population—for example, if all students leave school when they turn 16—only a very small control group would be needed to conclude confidently that students' increased educational attainment resulted from a successful program rather than from chance. The larger the underlying variation across people in the key outcome, the larger the treatment and control groups need to be. To continue the example, assume that equal numbers of students between the ages of 14 and 18 leave school every year and that the difference in the average age at the time they leave school is 16 for the controls and 17 for the participants. Given the same size sample as in the first example, it would be harder to be sure that the schooling difference was not attributable to randomly selecting more of the older leavers into the treatment group than into the control group.

Similarly, if the effect of a program is expected to be quite small, larger groups would be needed to be confident that any observed difference is not due to chance. For example, if a particular water treatment procedure was supposed to decrease the mortality rate by 1 percent, large groups of treatments and controls would be needed to conclude confidently that such a decrease was caused by the program; if the procedure was expected to decrease the mortality rate by 20 percentage points, much smaller groups of individuals would suffice.

The Long-Term Care Channeling demonstration (Kemper and others 1986) illustrates the relation between the expected effect and the size of the groups needed for a random assignment evaluation. This program attempted to enable the fragile elderly to stay in the community rather than have to go to a nursing home. It was expected that without the program 50 percent of the target population would go into a nursing home; policymakers wanted the program to reduce this rate to 45 percent. To be confident that a difference of 5 percentage points was not due to chance, evaluators calculated that the treatment and control groups would each have to consist of 1,715 individuals. The budget for the evaluation did not allow for such a large sample. Because they could fund only 1,000 individuals in each group, evaluators calculated that the rate would have to drop by 6.5 percentage points, to 43.5 percent, for them to conclude with confidence that the change was due to the program rather than to chance.

In sum, programs affecting outcomes that are inherently less variable or that will have larger impacts require smaller samples. It is, nonetheless, rare for randomized social program evaluations to consist of fewer than 500 to 1,000 participants in each group.

From this discussion, it should be clear that random assignment designs are not always feasible. If the number of group members needed to assess the impact is large and the potential recipients of the program are few, random assignment may not be a good option. For some types of interventions, such as programs aimed at entire communities, random assignment evaluation would be too costly. In these cases, careful judgmental matching of communities is a more reasonable way to construct a comparison group, even though conclusions based on such comparisons would be less robust than those based on randomized groups.

Operational Details

Details of a program can and should affect the choice of the evaluation method and how it is carried out. In particular, the evaluation should be designed to minimize any risk that the study itself might compromise the program—by altering the program's delivery in some fundamental way, changing the type of individual who would be served, or changing the behavior of the members of the counterfactual group.

ALTERING THE SERVICES DELIVERED. Sometimes, gathering information can itself be construed as a service. An important feature of the Long-Term Care Channeling program was an assessment by a clinician of the client's physical health and well-being. Much of the baseline information needed to learn whom the program would serve and for whom it was most effective was also medical data. How could such information be collected from the control group members without inadvertently providing them with part of the program? It was decided that interviews of the controls by nonmedical personnel would not

constitute an assessment. But if interviewers (not clinicians) were to gather the baseline information for the control group, perhaps interviewers should also question participants to ensure comparable data. A medical interview with an interviewer is not the same thing as a medical interview with a clinician, however, so this option was discarded. Similarly it was decided not to subject the participants to both the clinician's and the interviewer's questioning because this intense medical questioning would be exhausting for these fragile elderly and because it might be construed as a more intensive service. Ultimately, evaluators opted to live with noncomparably collected data—participants were interviewed by clinicians, controls were questioned by interviewers.

Less obviously, an evaluation can affect a program by overburdening program staff with research requirements. Evaluators should work closely with program staff to minimize the research burden on staff members. This is especially true for randomization schemes. A good example of how random assignment can be incorporated into a complex selection process without changing the essential elements of the program can be seen in the evaluation of the Big Brother/Big Sister program, in which adult volunteers serve as mentors and friends to children who have only one parent.

An essential element of the program is the careful matching of the adult with the child by trained social workers. Two alternative random assignment options were presented to the program personnel. In the first option, caseworkers would select two young people who they deemed appropriate for a particular volunteer. Then "a flip of the coin" would decide which youngster was actually assigned to the volunteer. The nonassigned child would be part of the control group. The second option randomly assigned new applicants to either the treatment or the control group. Children in the treatment group would then be bumped to the top of the caseworkers' lists to be matched as soon as possible, instead of having to wait the usual amount of time. The percentage of the treatment group that ultimately received matches would be higher in the first option than in the second, but the caseworkers' workloads would be twice what they were in the second. Both options preserved the matching element of the program. The design options were discussed with program staff, whose concerns were important in deciding which process was chosen. Finally, researchers decided to randomize the pool of new applicants before matching and so lessen the workload of an already overburdened staff. Then, before the process began, meetings were held at each site to explain and discuss the study and the randomization procedures and answer the staff's questions.

Researchers should make the randomization process tolerable and comprehensible to deliverers of the service because otherwise program staff may undermine the evaluation. Often, they are already overwhelmed with program requirements and resent the additional demands of the evaluation, which either entail more paperwork or oblige them to do their jobs slightly differently (for instance, to select two children per volunteer rather than one). Making the staff

part of the design process, providing funds to hire additional staff, or offsetting the extra work entailed can pay off handsomely in the long run.

Researchers should also be sensitive to program operators' need to serve a certain number of special cases (either for political reasons or because the applicant faces particular hardship). These exceptions can be made without seriously compromising the study, and they often buy evaluators the good will of the operators by indicating that the evaluation process is incorporating their point of view. A small number of participants can be allowed to circumvent the randomization process, by receiving services but being excluded from the research. However, program operators should be aware that evaluation results do not apply to these individuals. In sum, the process of randomization must be tailored to some degree to the needs of the operators.

CHANGING THE TARGET POPULATION. The research requirements of a demonstration can skew the selection of participants so that the population served by the pilot program differs from the intended recipients of the program itself. One reason is that individuals who are willing to participate in research may differ from individuals who are simply willing to receive services. For example, if extensive academic testing is needed to evaluate a youth education program, youngsters who might have participated but who are averse to being tested may fail to apply for the evaluated program. Similarly, if extensive recruiting is required to identify enough individuals to fill both the counterfactual and the treatment groups, the recruited individuals may differ from individuals who would come forward normally. This risk is minimized when there are more eligible individuals than the program can serve, so that no extra recruitment is needed.

Making sure that the target population for the demonstration is as similar as possible to that for the program is vital because the evaluation will assess the effect of the program only on those who were served. For example, the Long Term Care demonstration told policymakers how case management affected the fragile elderly but not the healthy elderly.⁴

CONTAMINATING THE COUNTERFACTUAL GROUP. When designing an evaluation strategy, researchers must consider how the nontreatment group will react to the presence of the evaluation. In many quasi-experimental designs—for example, reflexive or matched site designs—the individuals in the comparison group are completely unaware of the assessment. When other designs are used, however, especially randomized selection, the nontreatment group is aware of the experiment and may be adversely affected by this information. Individuals who are randomly selected into a control group, for example, may feel that they were passed over for treatment because of a personal failing, and this sense of rejection could cause them to act differently than they would have otherwise. This behavior inaccurately reflects how participants would have behaved had the program not existed; in other words, the counterfactual group

is contaminated. Similarly, enriching the educational opportunities of one class in a school might well dishearten the unselected students, whose performance might suffer as a result. The changed performance would compromise the effectiveness of these students as members of a comparison group for the evaluation.

A variant of this source of contamination is changed behavior of others in response to the selection process. If, for example, half the classes in a school were selected to receive additional resources, teachers might redirect their time to compensate nonparticipants for their lost opportunity. The resulting changed situation for the nonparticipants would damage their validity as comparisons.

A third source of contamination is the counterfactual group's unintended receipt of program services. That occurred in the early random assignment evaluation of the children's television program "Sesame Street," where some of the control children were found to have seen the program, thus compromising their value as a nontreatment group.

Program staff can often inform evaluators about potential contamination. In the Channeling demonstration program, for example, operators pointed out that people living together could not be randomized into different groups—the wife could not be a treatment while the husband was a control. For one thing, some of the program's services, such as housekeeping, were received by the entire household, so that the control would be receiving some program services. For another, envy could affect the control's behavior. This valid concern was addressed by randomizing households: when individuals applied to the program, they provided the names of everyone they lived with, and if someone they lived with had been randomized earlier, the new applicant received whatever designation—treatment or control—the previous household member had received.

Constraints on Time and Resources

Time and budget constraints also affect the choice of an evaluation strategy. All evaluation takes time, but random assignment and prospective comparison group designs (including prospective reflexive studies) take longer than retrospective comparison group designs. In particular, comparison group designs that select currently enrolled participants or even program graduates and compare them with other similar youth do not have to allow time for building up a sample for treatment and for assessing postprogram effects. Typically, the sample buildup period lasts a year. Then time must be allocated to allow the last member of the treatment group enrolled to receive the intervention. Finally, time is usually allocated to investigate whether the participants reap post-program benefits. There is always a tension between wanting results quickly and wanting to investigate long-term effects. Generally, the behavior of the two groups is examined twelve to eighteen months after sample enrollment.

Prospective studies are also often more costly than retrospective studies, as long as data for the retrospective studies are easily accessible. Prospective studies usually entail telephone or in-person interviews. Baseline (preprogram) interviewing is cheaper than follow-up interviews because one knows where the respondents are; for follow-up interviews, respondents must be tracked down. A common mistake of inexperienced evaluators is that they do not make every effort to find respondents and persuade them to answer the research questions. Attrition of the sample through nonresponse can seriously compromise the generalizability of the results. Sample attrition damages both randomized and comparison group designs. Generally, the evaluation community heavily discounts results from experiments in which the response rates are less than 70 percent. Unless the follow-up period is very long (four or five years or more), good-quality social evaluations usually are expected to have response rates of 80 to 85 percent.

Conclusions

The general consensus at present is that random assignment is the evaluation technique that produces the most defensible results. Ashenfelter and Card (1985, p. 648) conclude that “randomized clinical trials are necessary to determine program effects”; Barnow (1987, p. 190) says that “experiments appear to be the only method available at this time to overcome the limitations of non-experimental evaluations,” and Manski and Garfinkel (1992) labeled random assignment the “new orthodoxy.” But random assignment is not always feasible or suitable, and when that is so, evaluators and policymakers must settle for techniques that are inherently more open to criticism. Because these quasi-experimental methods rely extensively on statistical assumptions (Cook and Campbell 1979), explicit efforts should be made to collect the data needed to test those assumptions rigorously. (Moffitt 1991 provides a nontechnical discussion of the various estimation techniques and the tests appropriate for use with quasi-experimental data.)

But the greater credibility of random assignment should not lull researchers into a false sense of security and lax methodology. Random assignment in its purest form (comparison of simple means) works only if there is no systematic sample attrition and the control group does not become contaminated. Thus, data must be collected and statistical methods need to be employed to adjust for any of these problems that may arise.

Researchers should be careful in generalizing from experiments in general. Strictly speaking, estimated impacts describe only the effect that a particular program intervention, as delivered, had on the specific participants served. In drawing conclusions about the likely effectiveness of a program, it is vital to be clear about what intervention was actually delivered and thus what intervention was actually tested. The program “on the ground” could be quite

different from the program described in the rhetoric of the policymakers. For example, the Supported Work program was developed to move individuals gradually from a very structured job environment with a lot of supervision and relatively lax work standards to one in which the participant was doing the job more or less independently and meeting standards that would be demanded in unsubsidized jobs. But because “graduated stress” was never actually documented and related to outcomes, it would be incorrect to generalize from the Supported Work demonstration that graduated stress does or does not work for youth.

To be able to generalize findings adequately to other settings, one must understand the mechanisms by which the services changed the participants’ situation compared with that of members of the counterfactual group, and the means by which these services changed the participants’ decision process. *Why* did a program work or not work? Additional analysis should be conducted to explore the mechanisms through which the program had its effect. Heckman (1991) warns that it is all too easy for evaluators of randomized experiments to treat the program as a black box and fail to investigate in detail the social mechanisms by which it operates, because aggregate pronouncements of “effectiveness” or “ineffectiveness” can be made with relative ease.

Knowledge of what works is built up slowly. An evaluation rarely answers all questions. Yet, by understanding how public policy changed behavior, policymakers may be able to improve future interventions.

Notes

Jean Grossman is Director of Research and Vice President at Public/Private Ventures in Philadelphia, Pennsylvania.

1. Several textbooks discuss these issues further; Cook and Campbell (1979) and Rossi and Freeman (1989) are but two examples.

2. Some sample loss may be unavoidable (or even desirable). For example, if one of the goals of a program is to extend the lives of elderly individuals, it might well turn out that living members of the treatment group are, on average, more impaired than living members of the control group because more of the most impaired controls die, while impaired members of the treatment group have a higher probability of living. A simple comparison of the impairment levels of the two groups—as is usual in random assignment evaluations—would thus be inappropriate.

3. Full-coverage programs can be assessed by the experimental method only if a government has the will to waive the full-coverage requirement for the period of the research, as the U.S. government did to permit a random assignment strategy for the JTPA evaluation described earlier.

4. Effects for different populations can sometimes be inferred statistically if the alternative population does not differ too much from the original population (see Conlisk 1973).

References

The word “processed” describes informally reproduced works that may not be commonly available through library systems.

- Ashenfelter, Orley, and David Card. 1985. "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs." *Review of Economics and Statistics* 67:648–60.
- Barnow, Burt. 1987. "The Impact of CETA Programs on Earnings: A Review of the Literature." *Journal of Human Resources* 22:157–93.
- Bassi, Laurie J., Margaret C. Simms, Lynn C. Burbridge, and Charles L. Detsey. 1984. *Measuring the Effect of CETA on Youth and the Economically Disadvantaged*. Washington, D.C.: Urban Institute.
- Bloom, Howard, Larry Orr, George Cave, Stephan Bell, and Fred Doolittle. 1992. *The National JTPA Study—Title II-A Impacts on Earnings and Employment at 18 Month*. Bethesda, Md.: Abt Associates.
- Brewster, J. Alan, Walter Corson, John Friedmann, Walter Nicholson, and Andrea Vayda. 1978. *Follow-Up Study of Recipients of Federal Supplemental Benefits*. Princeton, N.J.: Mathematica Policy Research, Inc.
- Brown, Randall, John Burghardt, Edward Cavin, and Rebecca Maynard. 1984. "The Employment Opportunity Pilot Projects: Analytic Strategies and Estimates of Program Impacts." Princeton, N.J.: Mathematica Policy Research, Inc.
- Conlisk, John. 1973. "Choice of Response Function Form in Designing Subsidy Experiments." *Econometrica* 41(4):643–56.
- Cook, Thomas, and Donald Campbell. 1979. *Quasi-Experimentation Design and Analysis Issues for Field Settings*. Skokie, Ill.: Rand-McNally.
- Deniston, O., and I. Rosenstock. 1972. "The Validity of Designs for Evaluating Health Services." Research report. University of Michigan, School of Public Health, Ann Arbor. Processed.
- Devaney, Barbara, Linda Bilheimer, and Jennifer Shore. 1992. "Medicaid Costs and Birth Outcomes: The Effect of Prenatal WIC Participation and the Use of Prenatal Care." *Journal of Policy Analysis and Management* 11(4):573–92.
- Dickerson, Katherine P., Terry R. Johnson, and Richard W. West. 1984. *An Analysis of the Impact of CETA Programs on Participants' Earnings*. Menlo Park, Calif.: SRI International.
- Director, S. 1974. "Evaluating the Impact of Manpower Training Programs." Ph.D. diss., Northwestern University, Evanston, Ill. Processed.
- Farkas, George, Randell Olsen, Ernest Stromsdorfer, L. Sharpe, Felicity Skidmore, D. Smith, and S. Merrill. 1982. *Impacts from the Youth Incentive Entitlement Pilot Projects: Participation in Work and School over the Full Program Period*. New York: Manpower Demonstration Research Corporation.
- Fraker, Thomas, and Rebecca Maynard. 1984. "An Assessment of Alternative Comparison Group Methodologies for Evaluating Employment and Training Programs." Princeton, N.J.: Mathematica Policy Research, Inc.
- Friedlander, Daniel, and Philip K. Robins. 1992. "Estimating the Effects of Employment and Training Programs: An Assessment of Some Nonexperimental Techniques." Manpower Development and Research Corporation, New York.
- Heckman, James, J. 1980. "Sample Selection Bias as a Specification Error." In E. W. Stromsdorfer and G. Farkas, eds., *Evaluations Studies Review Annual*. Vol. 5. Beverly Hills, Calif.: Sage.
- . 1991. "Basic Knowledge—Not Black Box Evaluations." *Focus* 12(Summer): 234–35.
- . 1992. "Randomization and Social Policy Evaluation." In Charles Manski and Irwin Garfinkel, eds., *Evaluating Welfare and Training Programs*. Cambridge, Mass.: Harvard University Press.
- Heckman, James J., V. Joseph Hotz, and Marcelo Dabos. 1987. "Do We Need Experimental Data to Evaluate the Impact of Manpower Training On Earnings?" *Evaluation Review* 11(4):395–427.
- Heckman, James J., and Robert Robb. 1985a. "Alternative Methods for Evaluating the Impact of Interventions: An Overview." *Journal of Econometrics* 30:239–67.

- . 1985b. "Alternative Methods for Evaluating the Impact of Interventions: An Overview." In James J. Heckman and B. Singer, eds., *Longitudinal Analysis of Labor Market Data*. Cambridge, U.K.: Cambridge University Press.
- Hotz, V. Joseph. 1992. "Designing an Evaluation of the JTPA." In Charles Manski and Irwin Garfinkel, eds., *Evaluating Welfare and Training Programs*. Cambridge, Mass.: Harvard University Press.
- Job Training Longitudinal Survey Research Advisory Panel. 1985. "Recommendations of the Job Training Longitudinal Survey Research Panel." Prepared for the Office of Strategic Planning and Policy Development, Employment and Training Administration, U. S. Department of Labor, Washington, D.C. Processed.
- Kemper, Peter, George Carcagno, Randall Brown, Robert Applebaum, and Judith Wooldridge. 1986. *The Evaluation of the Long-Term Care Channeling Demonstration: Final Report*. Princeton, N.J.: Mathematica Policy Research, Inc.
- Kershaw, David, and Jennifer Fair. 1976. *The New Jersey Income-Maintenance Experiment*. New York: Academic Press.
- Lalonde, Robert. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Econometric Review* 76(4):604-20.
- Lalonde, Robert, and Rebecca Maynard. 1987. "How Precise Are Evaluations of Employment and Training Programs: Evidence from a Field Experiment." *Evaluation Review* 11(4):212-27.
- Maki, J. E., D. Hoffman, and Robert Berk. 1978. "A Time Series Analysis of the Impact of a Water Conservation Campaign." *Evaluation Quarterly* 2(February):107-18.
- Mallar, Charles, Stuart Kerachshy, Craig Thornton, and David Long. 1982. "Evaluation of the Economic Impact of the Job Corps." Princeton, N.J.: Mathematica Policy Research, Inc.
- Manski, Charles, and Irwin Garfinkel, eds. 1992. *Evaluating Welfare and Training Programs*. Cambridge, Mass.: Harvard University Press.
- Moffitt, Robert. 1991. "Program Evaluation with Nonexperimental Data." *Evaluation Review* 15(3, June):291-314.
- Rossi, Peter, and Howard Freeman. 1989. *Evaluation: A Systematic Approach*. Newbury Park, Calif.: Sage.
- Sipe, Cynthia, Jean Baldwin Grossman, and Julita Milliner. 1987. *The Summer Education and Training Program: A Report of the 1986 Experience*. Philadelphia, Pa.: Public/Private Ventures.
- Westat, Inc. 1984. *Summary of Net Impact Results of CETA*. Rockville, Md.
- Wolf, Wendy, Sally Liederman, and Richard Voith. 1987. *The California Conservation Corps: An Analysis of Short-term Impacts on Participants*. Philadelphia, Pa.: Public/Private Ventures.