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 drop-outs 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
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