KEY CONCEPTS FOR CHAPTER 8

Control group A group of untreated targets that is compared with experimental groups on outcome measures in impact evaluations.

Experimental group A group of targets to whom an intervention is delivered and whose outcome measures are compared with those of control groups.

Randomization Assignment of potential targets to experimental and control groups on the basis of chance.

CHAPTER 8

RANDOMIZED DESIGNS FOR IMPACT ASSESSMENT

This chapter describes and explains the use of randomized field experiments in impact assessments. Randomized experiments are based on comparisons between groups of targets randomly assigned either to experience some intervention or to be left "untreated." The randomized controlled experiment is the strongest research design for assessing net impacts of interventions. However, randomized experiments have their limitations, a key one being that they are applicable only to partial-coverage programs. Moreover, practical considerations of target and stakeholder cooperation, time, and costs, as well as considerations concerning human subjects, further limit their use. Nevertheless, even evaluators working in areas in which it is difficult to implement true experiments need to be familiar with them, because the logic of randomized controlled experiments is the basis for the design of all types of impact assessments and the analysis of the data from them.

This chapter provides an exposition of the basic ideas behind randomized experiments. Inasmuch as the logic of randomized experiments underlies all types of impact assessments, these ideas are important for all of the other impact assessment research designs.

UNITS OF ANALYSIS

At the outset, a note on units of analysis is important. Social programs may be designed to affect a wide variety of targets, including individuals, families, neighborhoods and communities, organizations such as schools and business firms, and political jurisdictions from counties to whole nations. In this and the next chapter, individual persons are generally used as examples of program targets to facilitate the exposition of important points. This usage should not be taken to imply that impact assessments are conducted only when persons are the intended intervention targets. The logic of impact assessment remains constant as one moves from one kind of unit to another, although the costs and difficulties of conducting field research may increase with the size and complexity of units. For instance, the confounding factors that affect individual students also influence students in classes; hence, whether one works with individual students or classes as the targets of an educational intervention, the same formal design considerations apply. Of course, the scale of field operations is considerably increased as one shifts from students to classes as targets. The samples in the two cases are composed of students and classes, respectively; however, gathering data on a sample of 200 student units is usually easier and considerably less costly than accumulating similar data on the same number of class units.

The choice of units of analysis is determined by the nature of the intervention and the target units to which it is delivered. A program designed to affect communities through block grants to local municipalities requires that the units studied be municipalities. An impact assessment of block grants that is conducted by contrasting two municipalities has a sample size of two—completely inadequate for many statistical purposes, even though observations may be made on large numbers of individuals within each of the two communities.

The evaluator attempting to design an impact assessment should begin by identifying the units designated as the targets of the intervention in question and hence to be specified as units of analysis. In most cases, defining the units of analysis presents no ambiguity; in other cases, the evaluator may need to carefully appraise the intentions of program designers. In still other cases, interventions may be addressed to more than one type of targets: A housing subsidy program, for example, may be designed to upgrade both the dwellings of individual poor families and the housing stocks of local communities. Here the evaluator may wish to design an impact assessment that consists of samples of individual households within samples of local communities, a design that is intended to estimate the net impacts of the program on individual households and also on the housing stocks of local communities.

EXPERIMENTS AS AN IMPACT ASSESSMENT STRATEGY

As we noted in Chapter 7, randomized experiments and other comparative designs can only be employed to assess the impacts of partial-coverage programs. Partial-coverage programs are those that either are to be tested on a trial basis or, for whatever reasons, are reaching only a portion of the eligible target population. Only in these circumstances is it possible to make appropriate comparisons between persons who are receiving the intervention and comparable persons who are not.

The Concept of Control and Experimental Groups

The net outcomes of an intervention can be conceptualized as the difference between persons who have participated in the program (the experimental group or groups) and *comparable* persons who have not (the control or comparison groups). If perfect comparability is achieved, the same extraneous confounding factors will be present in both groups; overall, both would be subject to the same degree to endogenous changes such as secular drift and the other extraneous confounding factors discussed in Chapter 7. Hence, if the two groups are perfectly comparable, the only differences between them will be caused by the intervention itself and by design effects, of which stochastic effects are the most important. Depending on the ways in which data are collected from the experimental and control groups, some design effects may be identical for the two groups. Stochastic effects, however, always cause differences to appear between the groups.

On the basis of the formula developed in the last chapter, a project's net effects can be expressed in terms of intervention and control groups as follows:

$$\begin{array}{c}\text{Net}\\\text{effect}\end{array} = \left[\begin{array}{c}\text{Gross}\\\text{outcome}\\\text{for an}\\\text{intervention}\\\text{group}\end{array}\right] - \left[\begin{array}{c}\text{Gross}\\\text{outcome}\\\text{for a}\\\text{comparable}\\\text{control group}\end{array}\right] \pm \left[\begin{array}{c}\text{Design}\\\text{effects and}\\\text{stochastic}\\\text{error}\end{array}\right]$$

A critical element in estimating net effects is identifying and selecting comparable experimental and control groups. Comparability between experimental and control groups means, in ideal terms, that the experimental and control groups are identical except for their participation or nonparticipation in the program under evaluation. More specifically, comparability requires the following:

Identical composition: Experimental and control groups contain the same mixes of persons or other units in terms of their program-related and outcome-related characteristics.

Identical predispositions: Experimental and control groups are equally disposed toward the project and equally likely, without intervention, to attain any given outcome status; in other words, selection effects should be identical in both groups.

Identical experiences: Over the time of observation, experimental and control groups experience the same time-related processes: maturation, secular drifts, interfering events, and so on.

Implementing Control Group Evaluations

Although perfect comparability could theoretically be achieved by matching each target in an experimental group with an identical target in a control group, this is clearly impossible in program evaluations. No two individuals, families, or other units are identical in all respects. An experimental biologist might attempt to achieve comparability in experiments by using animals from the same litter, but such matching is not possible for the evaluator, because even identical adult twins are not identical in their lifetime experiences and certainly are not available in sufficient numbers to be the sole basis for impact assessment.

Fortunately, one-to-one comparability is not necessary. It is only necessary for experimental and control groups to be identical in aggregate terms and in respects that are relevant to the intended effects of the program being evaluated. It may not matter at all for an impact evaluation that experimental and control group members differ in place of birth or vary slightly in age, as long as such differences do not influence the outcome variables. On the other hand, differences between experimental and control groups that are related to their assignment to their respective conditions are both relevant and especially important. Any characteristic that is related both to placement as an experimental or a control and to the intended outcome of the intervention can cause errors in estimates of net effect.

One of the important implications of these observations is that impact assessments require more than just a few cases. The larger the number of units studied (given the methods of selection we will discuss), the more likely experimental and control groups are to be statistically equivalent. Studies in which only one or a few units are in each group rarely, if ever, suffice for impact assessments because the odds are that any division of a small number of units will result in differences between them. (Important exceptions to this statement will be discussed in Chapter 9.)

For interventions that are likely to have small or variable effects, both experimental and control groups must be quite large. For example, in the Transitional Aid to Released Prisoners (TARP) experiments testing the impact of unemployment insurance eligibility on recidivism among ex-felons, the experimental groups contained close to 1,500 and the control groups nearly 2,500 ex-felons (Rossi, Berk, and Lenihan, 1980; see also Exhibits 8-E and 8-G). On the basis of previous evaluation evidence, it was expected that the intervention's effects would be small and quite variable from individual to individual so a large sample size was necessary to provide adequate statistical power. Conversely, if very large effects are produced by an intervention, they can be detected with a much smaller number of targets in the experimental and control groups.

Although we have so far characterized control groups as consisting of targets who receive no intervention, this is not always the case. More often, targets in control groups receive existing programs or alternative interventions. For example, the control group in an evaluation testing the effectiveness of a nutrition program may consist of persons who are following a variety of nutritional practices, some of their own devising and others directed by their doctors. What this means is that the impact of a program is estimated relative to whatever mix of interventions is experienced by the control targets.

There are basically four approaches to configuring comparable control and experimental groups. First, there is the randomized experiment method (Boruch, 1997), discussed in this chapter. Second, participants in programs may be contrasted with nonparticipants who have been selected for comparability in important respects—the nonrandomized comparison groups method. Third, participants may be compared with nonparticipants while controlling statistically for measured differences between participants and nonparticipants—the statistical controls method. Finally, one may pursue a mixed strategy in which randomized or nonrandomized controls and statistical controls are used together. The last three approaches are discussed in Chapter 9.

We should note that some evaluators distinguish between the terms *control group* and *comparison group*, the former denoting a group formed through random allocation of targets and the latter a group assembled nonrandomly. We do not follow this usage in most of this chapter because, in practice, the distinction is often only a matter of degree. Hence, we will use the term control group to refer to both control and comparison groups, except in discussions in which the distinction is important.

For the sake of convenience in exposition, the next few sections discuss randomized experimental designs in which only one intervention is being tested for impact. However, an important variation of the experimental design consists of comparing two or more programs in a systematic way. In this case, there may be several experimental groups—for example, a number of groups that are each following a particular nutritional regimen—with the net effects of each estimated relative to the others being tested. The designs to be discussed in this

chapter can easily be extended to involve the simultaneous testing of several alternative interventions (or combinations of interventions). Indeed, there is much to be gained in the way of useful information for policymakers and project managers if evaluations of several interventions are undertaken comparatively, so that a given intervention is compared not only to a control condition in which no intervention is made but also to alternative interventions. Multiple-intervention impact assessments provide more information on such issues as how best to modify interventions, alone or in combination, to maximize effects at a given level of funding. These complex designs are discussed in more detail later in the chapter.

Using Randomization to Establish Comparability

The best way to achieve comparability between experimental and control groups is to randomly allocate members of a target population to the two groups, allowing chance to decide whether a person (or other unit) is offered a program or is left untreated. It is important to note that "random" in this sense does not mean haphazard or capricious. On the contrary, randomly allocating targets to experimental and control groups requires taking extreme care to ensure that every unit in a target population has the same chance as any other to be selected for either group. This requires the application of some explicit chance-based procedure to the assignment process, for example, a random number table, roulette wheel, roll of dice, or the like.

Because the resulting experimental and control groups differ from one another only by chance, whatever influences may be competing with an intervention to produce outcomes are present in both groups to the same extent, except for chance fluctuations. For example, because of randomization, persons who would be more likely to seek out the program if it were offered to them on a free-choice basis are equally likely to be in the experimental as in the control group. Hence, both groups should have the same proportion of persons favorably predisposed to the intervention. The confounding factor of self-selection, therefore, cannot affect whatever outcome differences are observed between the groups because it has not influenced which group the targets are actually assigned to.

Of course, even though target units are assigned randomly, the experimental and control groups will never be exactly comparable. For example, by chance more women may end up in the control group than in the experimental group. But if the random assignment were made over and over, these fluctuations would average out to zero. The expected proportion of times that a difference of any given size on any given characteristic will be found in a long series of randomizations can be calculated from appropriate statistical models. Any given difference in outcome among randomized experimental and control groups, therefore, can be compared to what is expected on the basis of chance (i.e., the randomization process). Statistical testing thus lets a judgment be made as to whether a specific difference is likely to have occurred simply by chance or whether it is unlikely by chance and, hence, more likely represents the effect of the intervention. Because the intervention in a well-run experiment is the only difference other than chance between experimental and control groups, such judgments become the basis for discerning the existence of a net effect. The statistical procedures for making such calculations are quite straightforward and may be found in any text dealing with statistical inference.

Randomization Is Not Random Sampling

It is important not to confuse *randomization* (i.e., random assignment), in the sense used here, with *random sampling*. Whereas randomization means taking a set of units and assigning each unit to an experimental or control group by means of some randomizing procedure, random sampling consists of *selecting* units in an unbiased manner to form a representative sample from a population. Thus, researchers might use random sampling to select a representative group for study from a target population and then use random assignment to allocate each member of the sample to experimental or control conditions. Although the use of random samples to form a set of targets that is then randomized to form experimental and control groups is a highly recommended procedure, many randomized experiments are conducted using sets of targets that have not been selected by random sampling (i.e., that do not necessarily represent a given population). This latter procedure, of course, may not be a sensible course to follow because of the potential loss of generalizability.

Randomization Procedures

Randomization is technically easy to accomplish. Tables of random numbers are included in most elementary statistics or sampling textbooks. Larger tables of random numbers are also available in published form. Many computer statistical packages contain subroutines that generate random numbers. Even some of the better hand calculators have random-number generators built into them. Flipping coins or rolling (fair) dice are also effective ways of randomizing (see Boruch, 1997, and Boruch and Wothke, 1985, for discussions of how to implement randomization).

The Logic of Randomized Experiments

A typical randomized experimental design can be represented by the following modification of our basic impact assessment formula:

$$\text{Net effects} = \begin{bmatrix} \text{Scores on} \\ \text{postintervention} \\ \text{outcome} \\ \text{measures for} \\ \text{randomized} \\ \text{experimental} \\ \text{group} \end{bmatrix} - \begin{bmatrix} \text{Scores on} \\ \text{postintervention} \\ \text{outcome} \\ \text{measures for} \\ \text{randomized} \\ \text{control} \\ \text{(unexposed)} \\ \text{group} \end{bmatrix} \pm \begin{bmatrix} \text{Design} \\ \text{effects and} \\ \text{stochastic} \\ \text{error} \end{bmatrix}$$

Note that this representation assumes only postintervention measurement on outcome measures. Later in this chapter we consider what is to be gained or lost by employing after-only measures versus having measures before and after an intervention.

Exhibit 8-A presents a schematic view of a simple before-and-after randomized controlled experiment, indicating the logic behind the estimates of net effects that can be computed. Of course, the differences between the experimental and control groups, $E - C$, necessarily contain the stochastic effects described in Chapter 7. Hence, it would be necessary to apply tests of statistical inference to judge whether, in any particular case, the value of $E - C$ is likely to be due to stochastic error. Conventional statistical tests for before-and-after experiments include t tests, analysis of variance, and analysis of covariance (with the pretest as the covariate).

EXHIBIT 8-A Schematic Representation of a Randomized Experiment

	Outcome Measures		
	Before Program	*After Program*	*Difference*
Experimental group	E1	E2	E = E2 - E1
Control group	C1	C2	C = C2 - C1

Net effects of program = E - C, where
E1, C1 = measures of outcome variable before the program is instituted, for experimental and control groups, respectively
E2, C2 = measures of outcome variable after program is completed, for experimental and control groups, respectively
E, C = gross outcomes for experimental and control groups, respectively

Note that the schematic presentation in Exhibit 8-A defines effects as differences between before- and after-intervention measures of outcome. As we have mentioned earlier, for some types of outcomes, a preintervention measure is not possible to define. There are statistical advantages to having both before and after measures, however, and estimates of effects can be more precise when before measures are used to hold constant each individual target's starting point prior to the intervention. The critical measurements, of course, are the postintervention outcome measures for both experimentals and controls.

Examples of Randomized Experiments in Impact Assessment

Exhibit 8-B describes a randomized experiment to test the effectiveness of an intervention to change the poor eating habits of schoolchildren. Several of the experiment's features are relevant here. First, note that schools were the unit of analysis and, correspondingly, entire schools were assigned to either the experimental or control conditions. Second, note that a number of output measures were employed, covering the multiple nutritional objectives of the intervention. It is also important that statistical tests were used to aid in judging whether the net effects, in this case the experimental group's lower intake of overall calories and calories from fat, were simply a chance difference.

Exhibit 8-C describes a randomized experiment testing the effectiveness of case management provided by former psychiatric patients relative to that provided by the usual mental health personnel. This example illustrates the use of experimental design to compare the effects of a service innovation with customary service. It thus does not address the question of whether case management has effects relative to a control condition of no case management but, rather, evaluates whether a different approach would have better effects than current practice. Another interesting aspect of this impact assessment is the sample of clients who participated in the experiment. Although a representative group of clients eligible for case management was recruited, 25% declined to participate (which, of course, is their right), leaving some question as to whether the results of this experiment can be generalized to all eligible clients. This is rather typical of service

EXHIBIT 8-B CATCH: A Field Experiment on a Demonstration Program to Change the Dietary Habits of Schoolchildren

According to the Recommended Dietary Allowances, Americans on average consume too many calories derived from fats, especially unsaturated fats, and have diets too high in sodium. These dietary patterns are related to high incidences of coronary diseases and obesity. The Heart, Lung and Blood Institute, therefore, sponsored a randomized field experiment of an intervention designed to bring about better nutritional intake among school children, the Child and Adolescent Trial for Cardiovascular Health (CATCH).

CATCH was a randomized controlled field trial in which the basic units were 96 elementary schools in California, Louisiana, Minnesota, and Texas, with 56 randomly assigned to be intervention sites and 40 to be controls. The intervention program included training sessions for the food service staffs informing them of the rationale for nutritionally balanced school menus and providing recipes and menus that would achieve that goal. Training sessions on nutrition and exercise were given to teachers, and school administrations were persuaded to make changes in the physical education curriculum for students. In addition, efforts were made to reach the parents of participating students with nutritional information.

An analysis of the lunches served in the intervention and control schools showed that by the end of the three-year trial, the total calories provided in lunch meals declined in the intervention schools whereas there was a slight increase in the control schools, with a statistically significant difference between the two. Similar statistically significant differences favoring intervention schools were found with respect to the percentage of calories obtained from total fat and saturated fat. On the downside, there were no decreases in the cholesterol or sodium content of meals served in the intervention schools.

Importantly, the researchers found that participation in the school lunch program did not decline in the intervention schools, nor was participation lower than in the control schools. At baseline the participation rates were 72% for the intervention schools and 74% for the control schools; at the end of the experiment the rates were 70% and 74%, respectively.

Measured by 24-hour dietary intake interviews with children at baseline and at the 1994 follow-up, children in the intervention schools were significantly lower than children in control schools in total food intake, calories derived from fat and saturated fat, but no different with respect to intake of cholesterol or sodium. Because these measures include all food over a 24-hour period, they demonstrate changes in food patterns in other meals as well as school lunches. On the negative side, there was no significant lowering of the cholesterol levels in the blood of the students in intervention schools.

The CATCH study is strong evidence that the nutritional content of school lunches can be changed by relatively modest interventions with food service personnel, bolstered by nutrition education for the children. Whether both are essential to achieve change unfortunately is unknown.

SOURCE: Adapted from R. V. Luepker, C. L. Perry, S. M. McKinlay, P. R. Nader, G. S. Parcel, E. J. Stone, L. S. Webber, J. P. Elder, H. A. Feldman, C. C. Johnson, S. H. Kelder, and M. Wu, "Outcomes of a Field Trial to Improve Children's Dietary Patterns and Physical Activity: The Child and Adolescent Trial for Cardiovascular Health (CATCH)," *Journal of the American Medical Association* 275 (March 1996): 768-776.

EXHIBIT 8-C Assessing the Incremental Effects of a Service Innovation

A community mental health center in Philadelphia customarily provides intensive case management to clients diagnosed with a major mental illness or having a significant treatment history. Case managers employ an assertive community treatment (ACT) model and assist clients with various problems and services including housing, rehabilitation, and social activities. The case management teams are composed of trained mental health personnel working under the direction of a case manager supervisor.

In light of recent trends toward consumer-delivered mental health services, that is, services provided by persons who have themselves been mentally ill and received treatment, the community mental health center became interested in the possibility that consumers might be more effective case managers than nonconsumers. Former patients might have a deeper understanding of mental illness because of their own experience and may establish a better empathic bond with patients, both of which could result in more appropriate service plans.

To investigate the effects of consumer case management relative to the mental health center's customary case management, a team of evaluators conducted a randomized field experiment. Initially, 128 eligible clients were recruited to participate in the study; 32 declined and the remaining 96 gave written consent and were randomly assigned to either the usual case management or the experimental team. The experimental team consisted of mental health service consumers operating as part of a local consumer-run advocacy and service organization.

Data were collected through interviews and standardized scales at baseline and one month and then one year after assignment to case management. The measures included social outcomes (housing, arrests, income, employment, social networks) and clinical outcomes (symptoms, level of functioning, hospitalizations, emergency room visits, medication attitudes and compliance, satisfaction with treatment, quality of life). The sample size and statistical analysis were planned to have sufficient statistical power to detect meaningful differences, with especial attention to the possibility that there would be no meaningful differences, which would be an important finding for a comparison of this sort. Of the 96 participants, 94 continued receiving services for the duration of study and 91 of them were located and interviewed at the one-year follow-up.

No statistically significant differences were found on any outcome measures except that the consumer case management team clients reported somewhat less satisfaction with treatment and less contact with their families. While these two unfavorable findings were judged to warrant further investigation, the evaluators concluded on the basis of the similarity in the major outcomes that mental health consumers were capable of being equally competent case managers as nonconsumers in this particular service model. Moreover, this approach would provide relevant employment opportunities for former psychiatric patients.

SOURCE: Adapted from Phyllis Solomon and Jeffrey Draine, "One-Year Outcomes of a Randomized Trial of Consumer Case Management," *Evaluation and Program Planning*, 1995, 18(2): 117-127.

settings—there are almost always a variety of reasons why some appropriate participants in an impact assessment cannot or will not be included. Even when included, of course, there may be other reasons why final outcome measures cannot be obtained. In the experiment described in Exhibit 8-C, the evaluators were fortunate that only 2 of 96 original participants were lost to the study because they failed to complete service and only 3 were lost because they could not be located at the one-year follow-up.

Exhibit 8-D describes one of the largest and best known field experiments relating to national policy ever conducted in the evaluation field. This was an experiment to determine whether providing income support payments to poor, intact (i.e., two-spouse) families would cause them to reduce the amount of their paid employment, that is, create a work disincentive. The study was the first of a series of five, each varying slightly from the others, run by the Office of Economic Opportunity and the Department of Health, Education and Welfare (later, its successor agency, the Department of Health and Human Services) to test various forms of guaranteed income and their effects on the work efforts of poor and near-poor persons. All five of the experiments were run over relatively long periods, the longest involving more than five years; all had difficulties maintaining the cooperation of the initial groups of families involved; and all found that the income payments created a slight work disincentive, especially for teenagers and mothers with young children—those in the secondary labor force (Mathematica Policy Research, 1983; Robins et al., 1980; Rossi and Lyall, 1976; SRI International, 1983).

Despite their power to sustain the most valid conclusions about the net effects of interventions, randomized experiments account for a relatively small proportion of impact assessments. Political and ethical considerations may rule out randomization, particularly when interventions simply cannot be withheld without violating ethical or legal rules (although the idea of experimentation does not preclude delivering some alternative intervention to a control group). Despite the obstacles to randomized evaluation designs, there is a clear consensus on their desirability for impact assessment (Cook and Campbell, 1979; Cronbach, 1982; Mohr, 1995) and a growing literature on how to enhance the chances of success (Boruch, 1997; Dennis, 1990; Dunford, 1990). Moreover, many examples of the application of experimental design to impact assessment, such as those cited in this chapter, demonstrate their feasibility under appropriate circumstances.

Nonetheless, randomized field experiments are challenging to implement; costly if done on a large scale; and demanding with regard to the time, expertise, and cooperation of participants and service providers that are required. They are thus generally conducted only when circumstances are especially favorable, for instance, when a scarce service can be allocated by a lottery or equally attractive program variations can be randomly assigned, or when the impact question has especial importance for policy. Dennis and Boruch (1989) identified five threshold conditions that should be met before a randomized field experiment is undertaken (summarized by Dennis, 1990):

- The present practice must need improvement.
- The efficacy of the proposed intervention must be uncertain under field conditions.

EXHIBIT 8-D The New Jersey-Pennsylvania Income Maintenance Experiment

In the late 1960s, when federal officials concerned with poverty began to consider shifting welfare policy to provide some sort of guaranteed annual income for all families, the Office of Economic Opportunity (OEO) launched a large-scale field experiment to test one of the crucial issues in such a program: the prediction of economic theory that such supplementary income payments to poor families would be a work disincentive.

The experiment was started in 1968 and carried on for three years, administered by Mathematica, Inc., a research firm in Princeton, New Jersey, and the Institute for Research on Poverty of the University of Wisconsin. The target population was intact families with income below 150% of the poverty level whose male heads were between 18 and 58. The eight experimental conditions consisted of various combinations of income guarantees, pegged to what was then the current "poverty level" and the rates at which payments were taxed (adjusted to earnings received by the families). For example, in one of the conditions a family received a guaranteed income of 125% of the then-current poverty level, if no one in the family had any earnings. If their plan then had a tax rate of 50% and someone in the family received earned income, their payments were reduced 50 cents for each dollar earned. Other conditions consisted of tax rates that ranged from 30% to 70% and guarantee levels that varied from 50% to 125% of the poverty line. A control group consisted of families who did not receive any payments.

The experiment was conducted in four communities in New Jersey and one in Pennsylvania. A large household survey was first undertaken to identify eligible families, then those families were invited to participate. If they agreed, the families were randomly allocated to one of the experimental groups or to the control group. Families in the experimental groups reported their earnings each month and, if eligible for transfer payments, a check was mailed to them.

The participating families were interviewed in great detail prior to enrollment in the program and at the end of each quarter over the three years of the experiment. Among other things, these interviews collected data on employment, earnings, consumption, health, and various social-psychological indicators. The researchers then analyzed the data along with the monthly earnings reports to determine whether those receiving payments diminished their work efforts (as measured in hours of work) in relation to the comparable families in the control groups.

Although about 1,300 families were initially recruited, by the end of the experiment 22% had discontinued their cooperation. Others had missed one or more interviews or had dropped out of the experiment for varying periods. Fewer than 700 remained for analysis of the continuous participants. The findings were that experimental group families decreased their work effort by about 5%.

SOURCE: Adapted from D. Kershaw and J. Fair, *The New Jersey Income-Maintenance Experiment,* vol. 1 (New York: Academic Press, 1976).

- There should be no simpler alternatives for evaluating the intervention.
- The results must be potentially important for policy.

- The design must be able to meet the ethical standards of both the researchers and the service providers.

Some of the conditions that facilitate or impede the use of randomized experiments to assess impact are discussed in a later section of this chapter.

Near Experiments: Conditions of "Ignorability"

The desirable feature of randomization is that it is a sure way of achieving unbiased allocation of eligible targets to the experimental and control groups. Unbiased allocation requires that the probability of ending up in either the experimental or control group is identical for all participants in the study. Correspondingly, biased assignment occurs when individuals with certain characteristics have a higher probability than others of being selected for either group. In constituting experimental and control groups from a population with equal proportions of men and women, for example, an assignment procedure would be biased if members of one sex or the other were more likely to be in either the experimental or the control group.

There are several alternative ways of obtaining experimental and control groups that are close to those resulting from randomization, although each has some drawbacks. In addition, there are conditions under which it can be argued that groups have differences, but that such differences can be ignored as potential producers of bias.

Perhaps the most commonly used substitute for randomization is systematic assignment from serialized lists, a procedure that can often accomplish the same end as randomization, provided that the lists are not ordered in some way that results in a bias. For example, in allocating high school students to experimental and control groups, it might be sensible to place all those with odd ID numbers into the experimental group and all those with even ID numbers into a control group. As long as odd and even numbers were not originally assigned to differentiate among students according to some characteristic, the result will be statistically the same as random assignment. Of course, if the school has given odd ID numbers to female students and even numbers to males, this systematic bias would create experimental and control groups that each contained only one sex. Before using such systematic selection procedures, therefore, researchers must establish how the agency that generated the list accomplished serialization and judge whether the numbering process might produce unwanted systematic differences between various sections of the list.

Sometimes ordered lists of targets have subtle biases that are difficult to detect. For example, an alphabetized list might tempt one to assign, say, all persons whose last names begin with D to the experimental group and those whose last names begin with H to the control group. In a New England city, this would result in an ethnically biased selection—many names of French Canadian origin begin with D (e.g., DeFleur), whereas very few Hispanic names begin with H. Similarly, numbered lists may contain age biases if numbers are assigned sequentially. The federal government assigns Social Security numbers sequentially, for instance, so that individuals with lower numbers are generally older than those with higher numbers.

There are also circumstances in which biased allocation may be "ignorable" (Rog, 1994; Rosenbaum and Rubin, 1983). Occasionally, unplanned interventions occur in situations that can be regarded as unbiased and hence equivalent to a randomized experiment. An example from a study of flood effects on the growth of housing and population stocks illustrates a plausibly valid "natural" experiment. Hydrologic engineers have marked off the flood plains of most American rivers into regions characterized by the expected frequency of floods. Thus, the "ten-year flood plain" includes those regions in a river basin in which floods are expected to occur, on average, once in every decade. Although each year the areas within the ten-year flood plain have a one-in-ten chance of experiencing a flood, whether or not a flood occurs in a particular year in a particular spot can be regarded as a random event. Neighborhoods built on flood plains can thus be divided into "experimentals" (those in which floods actually occurred during, say, a two-year period) and "controls" (those in which no floods occurred). Because both sets of neighborhoods had the same probability of experiencing floods, they constitute natural experimental and control groups. Growth trends in the two groups can then be compared to estimate the impact of floods on the growth of housing stocks and population.

Of course, floods are events that can be understood as the effects of known natural processes and thus are not truly random events. But because those processes do not "select" some particular flood plains more than others, floods may be regarded for our purposes as virtually random events. In addition, our knowledge of the processes that create floods gives no indication that the kinds of housing and population located in the flood plains affect the chances of floods occurring in those places over any given period of time. Note, however, that the validity of this approach depends heavily on whether the hydrologists have correctly

marked out the ten-year flood plain. Because such maps are made partly on the basis of historical experience and partly on the basis of knowledge about the behavior of rivers in various terrains, the flood plain contours are subject to error.

Another circumstance frequently encountered involves using overcapacity targets as controls. For example, in a Minneapolis test of the effectiveness of a program to keep children in their families who might be placed in foster care, those children were placed in a no-intervention control group who could not be served by the family counseling program because the counseling agency was at full capacity at the time of referral (AuClaire and Schwartz, 1986). The "ignorability" assumption made was that when a child was referred had little or nothing to do with the child's prospects for reconciliation with his or her family.

Whether natural or unplanned events in fact are unbiased or have biases that can be safely ignored must be judged with close scrutiny of the circumstances of those events. Indeed, most circumstances that often are called "natural experiments" cannot be regarded as such in the strict sense of the term. If there is any reason to suspect that the events in question were likely to affect units (persons, communities, etc.) with certain characteristics more than others, then the conditions for a virtual experiment do not exist unless those characteristics can be confidently declared irrelevant to the intervention and outcomes to be studied. For example, communities that have fluoridated their water supplies cannot be regarded as an experimental group to be contrasted with those who have not, because communities that adopt fluoridation likely have distinctive characteristics, for example, lower average age and more progressive government, that cannot be regarded as irrelevant and represent bias in the sense used here. Similarly, families that have purchased townhouses cannot be regarded as appropriate controls for those who have purchased freestanding homes, because the very act of making such purchases is an indicator of other potential differences between the two groups.

Data Collection Strategies for Randomized Experiments

Under some conditions, the outcome variable can only be measured postintervention so that no pretest is possible. A program designed to help impoverished high school students go on to college, for instance, can be judged definitively only by whether experimentals go on to college more frequently than controls, a measure that can be taken only after the intervention. Such cases aside, the general rule is that the more measurements of the outcome variables made before and after the intervention, the better the estimates of net effects will be. Multiple longitudinal measurements increase measurement reliability and provide more information on which to build estimates of net outcomes. Measures taken before an intervention provide estimates of the preexperimental states of the experimental and control groups and are useful for making adjustments for preexisting differences between the two and for measuring how much of a gain the intervention effected. For example, preintervention measures of earnings for experimentals and controls in a vocational retraining project would enable researchers to make better estimates of the degree to which earnings improve as a result of training and at the same time would offer a variable to hold constant in the analysis of outcomes.

Periodic measurements taken during the course of an intervention are also useful. Such series allow evaluators to construct useful descriptive accounts of how an intervention works over time. For instance, if a vocational retraining effort is found to produce most of its effects during the first four weeks of a six-week program, this finding might lead to the suggestion that shortening the training period would cut costs without seriously curtailing the project's effectiveness. Likewise, multiple, periodic measurements can lead to a fuller understanding of how targets react to services. Some reactions may be slow-starting and then accelerate later; others may be strong initially but soon trail off to preintervention levels. For example, motorists' response to the 55-mile-per-hour speed limit is reputed to have consisted of an initial slowing down, followed by a gradual return to higher speeds. Being able to plot reactions to interventions allows evaluators to fine-tune programs for fuller effectiveness.

Thus, there are two compelling reasons for taking many measures before, during, and after an intervention. First, the more measures taken, the higher the reliability of composite measures. Second, interventions can be expected to have their effects over time; hence, longitudinal series can allow the evaluators to examine the way the intervention works over time.

ANALYZING RANDOMIZED EXPERIMENTS

Simple Randomized Experiments

The analysis of simple randomized experiments can be quite straightforward. Conducted properly, randomization produces experimental and control groups that are statistically equivalent. Hence, a comparison of outcomes in the two groups provides estimates of net effects. A comparison of these estimates, in turn, with the chance expectation derived from a statistical model then provides a means for judging whether those effects are larger than the chance fluctuations likely to appear when there really are no differences due to the intervention. Exhibit 8-E provides an example of the analysis conducted on a simple randomized experiment. The results are analyzed first by a simple comparison between experimentals and controls and then by means of a more complex multiple regression model.

Complex Randomized Experiments

It is common for impact assessment to involve tests of several variants of an intervention or several distinct interventions in a complex design. In the New Jersey-Pennsylvania Income Maintenance experiment (Exhibit 8-D), eight variations were tested, differing from one another in the amount of income guaranteed and the tax penalties on family earnings. These variations were included in the experiment to examine the extent to which work effort depended on the degree of work disincentive believed to be embodied in different payment schemes. A critical evaluation question was whether the work response to payments would vary with (a) the amount of payment offered and (b) the extent to which earnings from work reduced those payments.

Complex experiments along these lines are especially appropriate for testing new policies, because it may not be clear in advance exactly what form a new policy should or will take. A range of program variations provides more opportunity to cover the particular policy that might be adopted and hence increases the generalizability of the impact assessment. In addition, testing variations can provide information that helps guide program construction to optimize the effects and efficiency.

Exhibit 8-F, for example, describes a field experiment conducted on welfare policy in Minnesota. Two program variants were involved in the experimental conditions, both with more generous financial benefits to welfare clients who became employed and one with mandatory employment and training activities and one without. If these two versions of the program had proved equally effective, it would clearly be more cost-effective to implement the program without the mandatory employment and training activities and their associated administrative costs. However, the largest effects were found for the combination of financial benefits and mandatory training. This information allows policymakers to consider the trade-offs between the incrementally greater effects on income and employment of the more elaborate and expensive version of the program and the smaller, but still positive effects of the lower cost version of the program.

Under some circumstances, evaluators may be concerned that the administrative procedures proposed for a new program might compromise an otherwise effective intervention. The income maintenance experiments (Exhibit 8-D), for example, were criticized for requiring monthly income reports from each of the participating families (Rossi and Lyall, 1976). Because the welfare system ordinarily does not require such frequent reports from families receiving benefits, critics argued that this amounted to a stricter "means test" than that required by ordinary welfare regulations, and hence was potentially more demeaning. Where such concerns are serious, variations in the administrative procedures can be included in the experimental design to test their effects. Had the evaluators in the income maintenance

EXHIBIT 8-E Analysis of Randomized Experiments: The Baltimore LIFE Program

The Baltimore LIFE experiment was designed to test whether small amounts of financial aid to persons released from prison would help them make the transition to civilian life and reduce the probability of their being arrested and returned to prison. The financial aid was configured to simulate unemployment insurance payments, for which most prisoners are ineligible since they cannot accumulate work credits while imprisoned.

Persons released from Maryland state prisons to return to Baltimore were randomly assigned to either an experimental or control group. Those in the experimental group were told they were eligible for 13 weekly payments of $60 as long as they were unemployed. Those in the control group were told that they were participating in a research project but were not offered payment. Researchers periodically interviewed the participants and monitored the arrest records of the Baltimore Police Department for a year beyond each prisoner's release date. The arrest records yielded the results over the postrelease year shown in Table 8-E1.

TABLE 8-E1: Arrest Rates in the First Year After Release

Arrest Charge	*Experimental Group (n = 216)*	*Control Group (n = 216)*	*Difference*
Theft crimes (e.g., robbery, burglary, larceny)	22.2%	30.6%	-8.4
Other serious crimes (e.g., murder, rape, assault)	19.4%	16.2%	+3.2
Minor crimes (e.g., disorderly conduct, public drinking)	7.9%	10.2%	-2.3

The findings shown in the table are known as *main effects* and constitute the simplest representation of experimental results. Since randomization has made the experimental and control groups statistically equivalent except for the intervention, the arrest rate differences between them are assumed to be due only to the intervention plus any stochastic variability.

The substantive import of the findings is summarized in the last column on the right of the table, where the differences between the experimental and control groups in arrest rates are shown for various types of crimes. For theft crimes in the postrelease year the difference of -8.4 percentage points indicated a potential intervention effect in the desired direction. The issue then became whether 8.4 was within the range of expected chance differences, given the sample sizes (*n*). A variety of statistical tests are applicable to this situation, including chi-square, *t* tests, and analysis of variance. The researcher used a one-tailed *t* test, since the direction of the differences between the groups was given by the expected effects of the intervention. The results showed that a difference of -8.4 percentage points or larger would occur by chance less than five times in every hundred experiments of the same sample size (statistically significant at $p \leq .05$). The researchers concluded that the difference was large enough to be taken seriously as an indication that the intervention had its desired effect, at least for theft crimes.

The remaining types of crimes did not show differences large enough to survive the *t*-test criterion. In other words, the differences between the experimental and control groups were within the range where chance fluctuations were sufficient to explain them according to the conventional statistical standards ($p > .05$).

Given these results, the next question is a practical one: Are these differences large enough in a policy sense? In other words, would it be worthwhile to adopt the LIFE intervention as a social program? Would a reduction of 8.4 percentage points in theft crimes justify the payments and accompanying administrative costs? To answer this last question, the Department of Labor conducted a cost-benefit analysis (discussed in Chapter 11 in this volume) that showed that the benefits far outweighed the costs.

EXHIBIT 8-E Continued

A more complex and informative way of analyzing the theft crime data using multiple regression is shown in Table 8-E2. The question posed is exactly the same as in the previous analysis, but in addition, the multiple regression model takes into account the fact that many factors other than the payments might also affect arrests. The multiple regression analysis statistically controls those other factors while comparing the proportions arrested in the control and experimental groups.

TABLE 8-E2: Multiple Regression Analysis of Arrests for Theft Crimes

Independent Variable	*Regression Coefficient (b)*	*Standard Error of b*
Membership in experimental group	-.083*	.041
Unemployment rate when released	.041*	.022
Weeks worked the quarter after release	-.006	.005
Age at release	-.009*	.004
Age at first arrest	-.010*	.006
Prior theft arrests	.028*	.008
Race	.056	.064
Education	-.025	.022
Prior work experience	-.009	.008
Married	-.074	.065
Paroled	-.025	.051
Intercept	.263	.185
$R^2 = .094$*		
$N = 432$		

*Indicates significance at $p \leq .05$.

In effect, comparisons are made between experimentals and controls within each level of the other variables used in the analysis. For example, the unemployment rate in Baltimore fluctuated over the two years of the experiment: Some prisoners were released at times when it was easy to get jobs, whereas others were released at less fortunate times. Adding the unemployment rate at time of release to the analysis reduces the variation among individuals due to that factor and thereby purifies estimates of the intervention effect.

Note that all the variables added to the multiple regression analysis of Table 8-E2 were ones that were known from previous research to affect recidivism or chances of finding employment. The addition of these variables strengthened the findings considerably. Each coefficient indicates the change in the probability of postrelease arrest associated with each unit of the independent variable in question. Thus, the -.083 associated with being in the experimental group means that the intervention reduced the arrest rate for theft crimes by 8.3 percentage points. This corresponds closely to what was shown in Table 8-E1. However, because of the statistical control of the other variables in the analysis, the chance expectation of a coefficient that large or larger is much reduced

(continued)

EXHIBIT 8-E Continued

to only two times in every hundred experiments. Hence the multiple regression results provide more precise estimates of net effects. They also tell us that the unemployment rate at time of release, ages at release and first arrest, and prior theft arrests are factors that have a significant influence on the rate of arrest for these ex-prisoners and, hence, affect program outcome.

SOURCE: Adapted from P. H. Rossi, R. A. Berk, and K. J. Lenihan, *Money, Work and Crime: Some Experimental Evidence* New (York: Academic Press, 1980).

EXHIBIT 8-F Making Welfare Work and Work Pay: The Minnesota Family Investment Program

A frequent criticism of the Aid to Families With Dependent Children (AFDC) program is that it does not encourage recipients to leave the welfare rolls and seek employment because AFDC payments are typically more than could be earned in low-wage employment. The state of Minnesota received a waiver from the federal Department of Health and Human Services to conduct an experiment that would encourage AFDC clients to seek employment and allow them to receive greater income than AFDC would allow if they succeeded. The main modification embodied in the Minnesota Family Investment Program (MFIP) increased AFDC benefits by 20% if participants became employed and reduced their benefits by only one dollar for every three dollars earned through employment. A child care allowance was also provided so that those employed could obtain child care while working. This meant that AFDC recipients who became employed under this program had more income than they would have received under AFDC.

Over the period 1994 to 1996, some 15,000 AFDC recipients in a number of Minnesota counties were randomly assigned to one of three conditions: (1) An MFIP experimental group receiving more generous benefits and mandatory participation in employment and training activities; (2) an MFIP experimental group receiving only the more generous benefits and not the mandatory employment and training; and (3) a control group who continued to receive the old AFDC benefits and services. All three groups were monitored through administrative data and repeated surveys. The outcome measures included employment, earnings, and participation in education and training services.

An interim report covering 18 months and the first 9,000 participants in the experiment reported findings indicating that the demonstration was successful. MFIP experimental families were more likely to be employed and, when employed, had larger incomes than control families. Furthermore, those in the experimental group receiving both MFIP benefits and mandatory employment and training activities were more often employed and earned more than the experimental group receiving only the MFIP benefits.

SOURCE: Cynthia Miller, Virginia Knox, Patricia Auspos, Jo Anna Hunter-Manns, and Alan Orenstein, *Making Welfare Work and Work Pay: Implementation and 18-Month Impacts of the Minnesota Family Investment Program* (New York: Manpower Demonstration Research Corporation, 1997).

experiments configured an experimental group that operated under the ordinary income-reporting rules of the welfare system, the validity of the criticisms of the monthly reporting requirement could have been examined directly.

Of course, evaluators cannot endlessly proliferate experimental interventions to test every conceivable variation of a proposed program. For the income maintenance experiments, Kershaw and Fair (1976) proposed the concept of "policy space" as the basis for determining which program variations should be subject to testing. Policy space is the set of program alternatives that is likely to be politically acceptable if found effective and then considered by policymakers for implementation. Experimental assessment of innovative program concepts, in this view, should concentrate primarily on those variations that are clearly within the policy space defined by policymakers and administrators, perhaps extending a bit beyond, but not too far. In the income maintenance experiment, for instance, families of full-time students were excluded on the grounds that Congress would be very unlikely to make that group eligible, even though their income levels may have been well below the poverty line.

Analyzing Complex Experiments

As might be expected, complex randomized experiments require correspondingly complex modes of analysis. Although a simple analysis of variance may be sufficient to obtain an estimate of overall effects, the greater number of experimental groups and the amount of descriptive information typically collected on participants allow more elaborate forms of analysis. Sophisticated multivariate analysis, for instance, can provide greater precision in estimates of net effects and permit evaluators to pursue analytical themes not ordinarily available in simple randomized experiments. Exhibit 8-G provides an illustration of how a complex randomized experiment was analyzed through analysis of variance and causal modeling.

LIMITATIONS ON THE USE OF RANDOMIZED EXPERIMENTS

Randomized designs were initially formulated for laboratory and agricultural field research. Although their inherent logic is highly appropriate for the task of assessing the impact of social programs, they are nonetheless not applicable to all program situations. In this section, we review some of their limitations.

Programs in Early Stages of Implementation

As some of the examples in this chapter have shown, randomized experiments on demonstration programs can yield very useful information for purposes of policy and program design. However, once a program design has been adopted and implementation is under way, the impact questions randomized experiments are so good at answering are not usually appropriate to ask until the program is stable and operationally mature. In the early stages of program implementation, various features of a program often need to be changed for the sake of perfecting the intervention or its delivery. Although a randomized experiment can contrast program outcomes with those for untreated targets, the results will not be very informative if the program has changed during the course of the experiment. If the program has changed appreciably before outcomes are

EXHIBIT 8-G Analyzing a Complex Randomized Experiment: The TARP Study

Based on the encouraging findings of the Baltimore LIFE experiment described in Exhibit 8-E, the Department of Labor decided to embark on a large-scale experiment that would use existing agencies in two states to administer unemployment insurance payments to ex-felons. The objectives of the proposed new program were the same: Making ex-felons eligible for unemployment insurance was intended to reduce the need for them to engage in crime to obtain income. The payments in that sense were intended to compete with illegal activities as a source of income and to provide for income during a transition period from prison life to gainful employment.

The new set of experiments, called Transitional Aid to Released Prisoners (TARP), was also more differentiated in that it included varying periods of eligibility for benefits and varying rate schedules by which payments were reduced for every dollar earned in employment ("tax rates").

The main effects of the interventions are shown in the analyses of variance in Table 8-G1. (For the sake of simplicity, only results from the Texas TARP experiment are shown.) The interventions had no effect on property arrests: The experimental and control groups differed by no more than would be expected by chance. However, the interventions had a very strong effect on the number of weeks worked during the postrelease year: Ex-felons receiving payments worked fewer weeks on the average than those in the control groups and the differences were statistically significant. In short, it seems that the payments did not compete well with crime but competed quite successfully with employment!

Overall, these results seem to indicate that the experimental interventions did not work in the ways expected and indeed produced undesirable effects. However, an analysis of variance of this sort is only the beginning of the analysis. The results suggested to the evaluators that a set of counterbalancing processes may have been at work. It is well known from the criminological literature that unemployment for ex-felons is related to an increased probability that they will be rearrested and subsequently returned to prison. Hence, the researchers postulated that the unemployment benefits created a work disincentive represented in the fewer weeks worked by participants receiving more weeks of benefits or a lower "tax rate" and that this should have the effect of increasing criminal behavior. On the other hand, the payments should have reduced the need to engage in criminal behavior to produce income. Thus, a positive effect of payments in reducing criminal activity may have been offset by the negative effects of less employment over the period of the payments so that the total effect on arrests was virtually zero.

To examine the plausibility of this "counterbalancing effects" interpretation of the findings of the experiment, a causal model was constructed, as shown in Figure 8-G1. In that model, negative coefficients are expected for the effects of payments on employment (the work disincentive) and for their effects on arrests (the expected intervention effect). The counterbalancing effect of unemployment, in turn, should show up as a negative coefficient between employment and arrest, indicating that fewer weeks of employment are associated with more arrests. The coefficients shown in Figure 8-G1 were derived empirically from the data using a statistical technique known as three-stage least squares (more generally, structural equation modeling). As shown there, the hypothesized relationships appear in both the Texas and Georgia data.

EXHIBIT 8-G Continued

TABLE 8-G1: Analysis of Variance of Property-Related Arrests (Texas data)

A. Property-related arrests during postrelease year

Experimental Group		*Mean Number of Arrests*	*Percent Arrested*	*n*
26 weeks payment, 100% tax		.27	22.3	176
13 weeks payment, 25% tax		.43	27.5	200
13 weeks payment, 100% tax		.30	23.5	200
No payments, job placement[a]		.30	20.0	200
Interviewed controls		.33	22.0	200
Uninterviewed controls[b]		.33	23.2	1,000
ANOVA *F* value	=	1.15	.70	
p value	=	.33	.63	

B. Weeks worked during postrelease year

Experimental Group		*Average Number of Weeks Worked*	*n*
26 weeks payment, 100% tax		20.8	169
13 weeks payment, 25% tax		24.6	181
13 weeks payment, 100% tax		27.1	191
No payments, job placement		29.3	197
Interviewed controls		28.3	189
ANOVA *F* value	=	6.98	
p value	=	< .0001	

a. Ex-felons in this intervention group were offered special job placement services (which few took) and some help in buying tools or uniforms if required for jobs. Few payments were made.

b. Control observations made through arrest records only; hence no information on weeks worked.

This complex experiment, combined with sophisticated multivariate analysis, therefore, shows that the net effects of the intervention were negligible but also provides some explanation of that result. In particular, the evidence indicates that the payments functioned as expected to reduce criminal behavior but that a successful program would have to find a way to counteract the accompanying work disincentive with its negative effects.

(continued)

EXHIBIT 8-G Continued

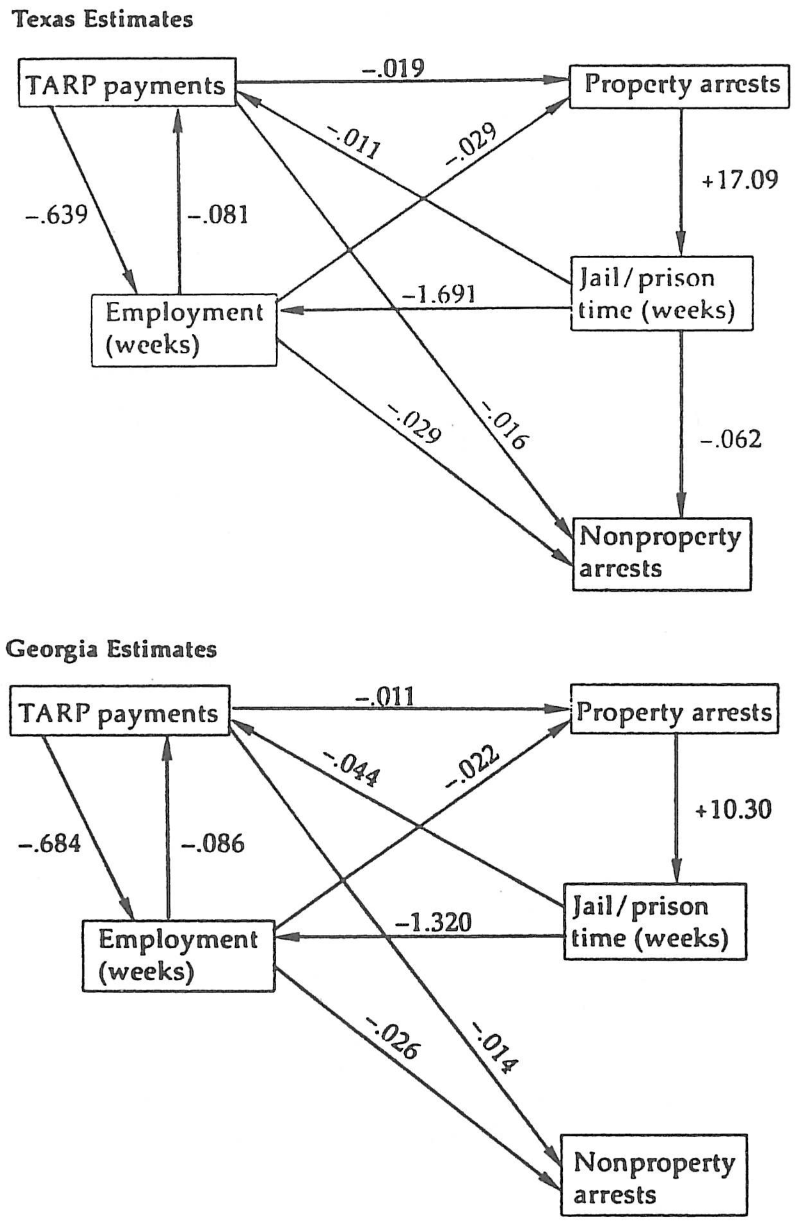

Figure 8-G1

SOURCE: Adapted from P. H. Rossi, R. A. Berk, and K. J. Lenihan, *Money, Work and Crime: Some Experimental Evidence* (New York: Academic Press, 1980).

measured on the participants, the effects of the different variants of the intervention are all mixed together in the experimental results, and there is no easy way to determine what effects are produced by any given form of the intervention.

For example, in a program that starts out providing group therapy but, in refining its services, ends up giving individual counseling, the overall results of an experiment would only indicate the effects of being in the program group relative to the control group. Because clients would not be randomly assigned to group versus individual counseling, that comparison would be contaminated by selection bias, and it would be difficult to determine the relative effectiveness of the different versions of the intervention. Moreover, if incremental program changes were adopted gradually over a period of time, it might not even be possible to establish just what conditions each participant experienced. Expensive field experiments, therefore, are best reserved for tests of firmly designed interventions that will be consistently implemented during the course of the experiment. An example of this strategy is a California experiment testing the effectiveness of a family preservation program designed to avert placement of abused or neglected children in foster or institutional homes by providing intensive services to the children and their families. A full-scale randomized experiment was started only after the agencies involved had two years of experience with running the program (Yuan, 1990).

Ethical Considerations

A frequent obstacle to the use of randomized experiments is that some stakeholders have ethical qualms about randomization, seeing it as arbitrarily and capriciously depriving control groups of positive benefits. The reasoning of such critics generally runs as follows: If it is worth experimenting with a program (i.e., if the project seems likely to help targets), it is a positive harm to withhold potentially helpful services from those who need them. To do so is therefore unethical. The counterargument is obvious: Ordinarily, it is not known whether an intervention is effective; indeed, that is the reason for an experiment. Because researchers cannot know in advance whether an intervention will be helpful, they are not depriving the controls of something known to be beneficial.

Sometimes an intervention may present some possibility of positive harm, and decision-makers may be reluctant to authorize randomization on those grounds alone. In some utility pricing experiments, for instance, there was a good chance that household utility bills would increase in some of the experimental groups. The researchers countered this argument by promising experimental households that any such overages would be reimbursed after the study was over. Of course, this promise of reimbursement changes the character of the intervention, possibly fostering irresponsible usage of utilities.

The most compelling ethical objections generally involve the conditions of control groups. If conventional services are known to be effective for their problems, it would generally be unethical to withhold those services for the purposes of testing an alternative to conventional services. We would not, for instance, deprive schoolchildren of mathematics instruction so that they could constitute a control group in an experiment testing a new math curriculum. In such cases, however, the important question is not whether the new curriculum is better than no instruction but, rather, whether it is better than current practices. The appropriate experimental comparison, therefore, is between the new curriculum and the control condition of current instructional practice with no student going without credible instruction.

When program resources are scarce and fall well short of demand, random assignment to control conditions can present an especially difficult ethical dilemma. This procedure amounts to randomly selecting those relatively few eligible targets who will receive the program services. If the intervention cannot be given to all who qualify, it can be argued that randomization is an equitable method of deciding who is to get it, because all targets have an equal chance. And, indeed, if there is great uncertainty about the efficacy of the intervention, this may be quite acceptable. However, when service providers are convinced that the intervention is efficacious, as they often are despite the lack of experimental evidence, they may object strongly to allocating service by lot and insist that the neediest targets receive priority. As will be discussed in the next chapter, this is a situation to which the regression-discontinuity design is well adapted, although it may be very problematic for randomized designs.

Differences Between Experimental and Actual Intervention Delivery

A third limitation is that intervention delivery in experimental conditions may be different in critical ways from intervention delivery when the program is implemented. Many major, large-scale field experiments, for example, have used money payments as interventions (e.g., the income maintenance experiment described in Exhibit 8-D). With such standardized and easily delivered interventions, researchers can be relatively certain that the experimental intervention will be similar to that of a fully implemented program, because there are only a limited number of ways in which checks can be delivered. However, more labor-intensive, high-skill interventions (job placement services, counseling, teaching, etc.) are likely to be delivered with greater fidelity to the designers' intentions in a field experiment than when they are implemented as a program. Indeed, as we saw in Chapter 6, the very real danger that interventions will deteriorate in implementation is one of the principal reasons for monitoring programs.

This possibility argues for at least two rounds of experiments: a first round in which interventions are tested in their purest form, and a second round in which effective methods of service delivery through public agencies are tested and compared. The two stages of experiments in the Department of Labor's program to provide unemployment insurance benefits to released prisoners described in Exhibits 8-E and 8-G used this strategy. The first stage consisted of the small-scale experiment in Baltimore involving 432 prisoners released from the Maryland state prisons. The researchers selected the prisoners before release, provided them with payments, and observed their work and arrest patterns for a year. As may be recalled from Exhibit 8-E, the results showed a reduction in theft arrests over the postrelease period for experimental groups receiving unemployment insurance payments for 13 weeks.

The much larger second-stage experiment was undertaken in Georgia and Texas with 2,000 released prisoners in each state (Exhibit 8-G). In this experiment, payments were administered through the Employment Security Agencies in each of the states, and the tracking of the released prisoners over the postrelease year was accomplished jointly by the state prison systems and employment security agencies. The second-stage experiment was close to the system of administration that would have been put into place if the program had been enacted through federal legislation. The second-stage results, however, found the payments to be ineffective when administered under existing Employment Security Agency rules.

Time and Cost

An influential obstacle to the use of randomized experiments is that they are usually costly and time-consuming, especially large-scale multisite experiments. Ordinarily, they should not be undertaken to test program concepts that lie outside any conceivable policy space and so will never be considered, or to test established programs when there is not significant policy interest in evidence about impact. Moreover, experiments should not be undertaken when information is needed in a hurry. To underscore this last point, it should be noted that the New Jersey-Pennsylvania Income Maintenance experiment (Exhibit 8-D) cost $34 million (in 1968 dollars) and took more than seven years from design to published findings. The Seattle and Denver income maintenance experiments took even longer, with their results appearing in final form long after income maintenance as a policy had disappeared from the national agenda (Mathematica Policy Research, 1983; Office of Income Security, 1983; SRI International, 1983).

Generalizability and Validity

Because randomized experiments require such tight controls on interventions and the selection of participants, they are likely not to have very high generalizability or external validity. No field experiment evaluating a social program has ever been conducted using a sample of clients drawn from the entire population of the United States. The administrative complexities of running national experiments have seemed too severe a burden for the designers to attempt such a study. In practice, although randomized field experiments may vary in scale, they generally are best reserved for testing services that can be standardized and easily transferred to operating agencies, and for which a relatively small number of sites or locales can be evaluated with reasonable confidence of broader external validity.

Integrity of Experiments

Finally, we should note that the integrity of a randomized experiment is easily threatened. Although randomly formed experimental and control groups are "statistically equivalent" at the start of an evaluation, nonrandom processes may threaten their equivalence as the experiment progresses. Differential attrition may introduce differences between experimentals and controls. In the income maintenance experiments, for example, families in the experimental groups who received the less generous payment plans and families in the control groups were more likely to stop participating. Also, administrative procedures for arranging the intended intervention and control conditions may fail so that the comparison between them does not actually represent program effects (Exhibit 8-H provides an example of an experiment compromised in this way).

Also, it is difficult to deliver a "pure program." Although an evaluator may design an experiment to test the effects of a given intervention, everything that is done to the experi-

EXHIBIT 8-H A Compromised Impact Assessment: The New Jersey Family Cap Experiment

In the early 1990s, New Jersey asked for a waiver to AFDC rules to remove what was thought to be an incentive in the AFDC regulations that encouraged women to have additional children to increase their AFDC payments. New Jersey proposed to change its AFDC policy to establish a "family cap" prohibiting any payment increases for children conceived by an AFDC recipient after her first enrollment in the program. The Department of Health and Human Services agreed to the request but insisted that the effectiveness of the program be evaluated through a randomized experiment. The family cap went into effect in 1992 covering all AFDC families with the exceptions noted below. The new regulations were widely publicized in the state's mass media and were carefully explained to ongoing and newly enrolled AFDC participants.

A randomized experiment was designed in which some 6,000 AFDC families were randomly assigned to either an experimental group, whose additional children would not lead to payment increases, and a control group operating under the old AFDC rules in which grants were increased for children born while on AFDC ten or more months after enrollment. Case workers who were assigned to control group families were instructed to explain that the new family cap rules did not apply to them and letters were also sent to each family with that information. The evaluation plan was to track births and abortions occurring in the experimental and control groups through administrative data, including Medicaid records and periodic interviews. Comparisons between experimental and control families would then be used to determine if the family cap policy led to fewer additional births.

About two years into the experiment, the Rutgers researchers found that more than 20 families in the control group to whom a child had been born had been denied AFDC payment increases. In addition, a survey conducted of families in the experiment found that almost half of the women in the control group believed that their grants would not be increased if they had additional children, that is, they believed that they were subject to the family cap rules.

Apparently, the implementation of the research design had been compromised. Caseworkers failed to treat control families as intended and the control families did not understand that they were exempted from the family cap rules. Possibly the wide publicity given to the family cap simply overwhelmed whatever information was communicated to the control group families. It is also possible that not enough effort was made to communicate to caseworkers the special rules that applied and to ensure that participants knew about those rules. Most likely, both processes were at work.

Although a research team from Rutgers University attempted to estimate the impacts of the waiver, the apparent implementation failure of the experiment led some to question the credibility of the estimates.

SOURCE: M. J. Camasso, C. Harvey, and R. Jaganathan, *An Interim Report on the Impact of New Jersey's Family Development Program* (New Brunswick, NJ: Rutgers University School of Social Work, 1996).

mental targets becomes part of the intervention. For example, the TARP experiments (Exhibit 8-G) were supposed to test the effects of modest amounts of postprison financial aid, but the aid was administered by an existing state agency and hence that latter's procedures became part of the intervention. Indeed, there are few, if any, large-scale randomized social experiments that have not suffered some dilution. Of course, even if randomization is compromised to some extent, the results of a randomized experiment, properly analyzed, may still be superior in credibility to the nonrandomized designs discussed in the next chapter.

SUMMARY

- Randomized experiments are the flagships of evaluation. They generally provide the most credible conclusions about the impact of social programs. Policymakers, stakeholders, and the general public are most likely to treat findings emerging from true experiments respectfully, because they are familiar with at least the outlines of such designs from an awareness of the way laboratory studies are conducted.
- The designs and analysis procedures of all impact assessments are kin to those of true experiments; thus, an appreciation of experiments is important for anyone undertaking impact evaluations or using their results.
- The choice of units of analysis in impact assessments is determined by the nature of the intervention and the targets to which the intervention is directed.
- Randomized experimental designs are applicable only to partial-coverage programs in which there are sufficient untreated targets from which to draw a control or comparison group.
- The ideal experiment isolates the effect of the intervention being evaluated by ensuring that experimental and control groups are exactly comparable except for the intervention received. Strictly comparable groups are identical in composition, experiences over the period of observation, and predispositions toward the program under study. In practice, it is sufficient that the groups, as aggregates, are alike with respect to any characteristics that could be relevant to the intervention outcome.
- Randomization is a technique for ensuring comparability of experimental and control groups by distributing extraneous factors equally across the groups. Although stochastic effects will create some differences between any two groups, statistical procedures enable researchers to estimate the likelihood that observed differences are due to chance rather than to the intervention being studied.
- Assuming a well-run experiment, the estimate of an intervention's net effects can be expressed as the experimental group's score on a postintervention measure minus the control group's score, plus or minus stochastic effects.

- Surrogate procedures, such as existing target lists or naturally occurring events, can sometimes substitute for randomization so long as the resulting assignments to experimental and control groups are free of biases relevant to the intervention and the expected outcome.
- Although postintervention measures of outcome are critical in impact assessments, measures taken before and during an intervention, as well as repeated measurements afterward, increase measurement reliability and the precision of estimates of net effects and enable researchers to reconstruct how the intervention worked over time.
- Simple randomized experiments are analyzed by means of a comparison of the outcomes of the experimental and control groups, together with statistical procedures for determining whether any observed differences are likely to be due to chance variations.
- More complex research designs can compare a number of variations of an intervention and can be especially appropriate for testing new policies when the exact form of the intervention has not been firmly established. This type of design can also be used to study variations in the mode of intervention delivery.
- Despite their rigor, randomized experiments have several limitations when applied to social programs:

    1. They may not be useful in the early stages of program implementation when interventions may change in ways not allowed for in the experiment.
    2. Randomization is sometimes perceived by stakeholders as unfair and even unethical because of the differential intervention given to experimental and, especially, control groups.
    3. The way in which intervention is delivered in the experimental condition may not resemble intervention delivery in the implemented program.
    4. Experiments are costly and time-consuming.
    5. Because experiments require tight controls, the results may be low in generalizability and external validity.