KEY CONCEPTS FOR CHAPTER 7

Impact	The net effects of a program (see **net effects**).
Gross outcomes	The overall outcome subsequent to intervention, only part of which might actually be caused by the intervention.
Net effects	The effects of an intervention that can be attributed uniquely to it, that is, with the influence of confounding effects from other sources controlled or removed. Also called net outcomes and net impact.
Confounding factors	Extraneous variables resulting in observed effects that obscure or exaggerate the true effects of an intervention.
Selection bias	A confounding effect produced by preprogram differences between program participants and eligible targets who do not participate in the program.
Design effects	The influence of the research methods and procedures on the estimate of the net effects of a program.
Measurement validity	The extent to which a measure reflects the concept it is intended to measure.
Reliability	The extent to which scores obtained on a measure are reproducible in repeated administrations (provided that all relevant measurement conditions are the same).
Stochastic effects	Measurement fluctuations attributable to chance.
Proxy measure	A variable used to stand in for one that is difficult to measure directly.
Randomized experiment	An impact research design in which experimental and control groups are formed by random assignment.
Quasi-experiment	An impact research design in which "experimental" and "control" groups are formed by a procedure other than random assignment.
Reproducibility	The extent to which the findings of a study can be reproduced by other researchers in replications.
Generalizability	The extent to which an impact assessment's findings can be extrapolated to similar programs or from the program as tested to the program as implemented.

CHAPTER 7

STRATEGIES FOR IMPACT ASSESSMENT

The ultimate purpose of a social program is to ameliorate some social problem or improve some social condition. If the program theory is sound and the program plan is well implemented, those social benefits are expected to follow. Rarely are those benefits assured, however. Practical and conceptual shortcomings combined with the intractable nature of many social problems all too easily undermine the effectiveness of social programs.

Impact assessments are undertaken to find out whether interventions actually produce the intended effects. Such assessments cannot be made with certainty but only with varying degrees of plausibility. A general principle applies: The more rigorous the research design, the more plausible the resulting estimate of intervention effects.

The design of impact evaluations needs to take into account two competing pressures: On the one hand, evaluations should be undertaken with sufficient rigor that relatively firm conclusions can be reached; on the other hand, practical considerations of time, money, cooperation, and protection of participants limit the design options and methodological procedures that can be employed.

Ordinarily, evaluators assess the effects of social programs by comparing information about outcomes for participants and nonparticipants, by making repeated measurements on participants before and after intervention, or by other methods that attempt to achieve the equivalent of such comparisons. The basic aim of an impact assessment is to produce an estimate of the net effects of an intervention—that is, an estimate of the impact of the intervention uncontaminated by the influence of other processes and events that also may affect the behavior or conditions at which a program is directed. The strategies available for isolating the effects attributable to an intervention and estimating their magnitude are introduced in this chapter, together with issues surrounding their use.

Impact assessment can be relevant at many points throughout the life course of social programs. At the stage of policy and program formation, impact assessments of pilot demonstrations are sometimes commissioned to determine whether the proposed program

would have the intended effects. At the stage of program design, impact evaluations may be undertaken to test for the most effective ways to develop and integrate the various program elements. For example, the relative impact of different durations of service, of one type of practitioner versus another, and of providing follow-up services or not to targets are all issues that can be addressed through impact assessment.

When a new program is authorized, it is often started initially in a limited number of sites. Obviously, it is unwise to implement a new program widely without some knowledge of its effects. Impact assessments may be called for to show that the program has the expected effects before extending it to broader coverage. Furthermore, in many cases the sponsors of innovative programs, such as private foundations, implement programs on a limited scale with a view to promoting their adoption by government agencies if their effects can be demonstrated. Moreover, knowledge of program effects is critical to decisions about whether a particular initiative should be supported in preference to competing social action efforts.

Also, programs may be modified and refined to enhance effectiveness or to accommodate revised program goals. Sometimes the changes made are major and the assessments of the modified program resemble those of innovative programs. At other times, the modifications are modest "fine-tuning" efforts and the skeleton of the program remains fundamentally the same. In either case, the modifications can be subjected to impact assessments.

Finally, many established programs can be subjected to impact assessments, either continually or periodically. For example, the high costs of certain medical treatments make it essential that their efficacy be continually evaluated and compared with other means of dealing with the same medical problem. In other cases, long-established programs are evaluated at regular intervals either because of "sunset" legislation requiring demonstration of effectiveness if funding is to be renewed or as a means of defending the programs against attack by supporters of alternative interventions or other uses for the public funds involved.

KEY CONCEPTS IN IMPACT ASSESSMENT

All impact assessments are comparative (see Exhibit 7-A). Determining impact requires comparing, with as much rigor as is practicable, the conditions of targets who have experienced an intervention with those of equivalent targets who have experienced something else. There may be one or more groups of targets receiving "something else," and "something else" may mean receiving alternative services or simply going untreated. The "equivalent" targets for comparison may be selected in a variety of ways or comparisons may be made between information about the behavior or condition being examined and similar information from the same targets taken at an earlier time, or between measures of outcomes and conjectures about what would have occurred in the absence of the intervention.

The Experimental Model

Although there are many ways in which impact assessments can be conducted, the options available are not equal: Some characteristically produce more credible estimates of impact than others. The options also vary in cost and level of technical skill required. As in other matters, the better approaches to impact assessment generally require more skills and more time to complete, and they cost more.

In this and subsequent chapters, our discussion of the available options is rooted in the view that the optimal way to establish the effects caused by an intervention is a randomized field experiment. The laboratory model of such experiments is no doubt familiar. Subjects in laboratory experiments are randomly sorted into two or more groups. One group is designated the control group and receives no intervention or an innocuous one; the other group or groups, called the experimental group(s), are given the intervention(s) being tested. Outcomes are then observed for both the experimental and the control groups, with any differences being attributed to the experimental intervention.

This research model underlies impact evaluations as well, because such evaluations, like laboratory experiments, are efforts to establish whether certain effects are caused by the intervention. Sometimes impact evaluations closely follow the model of randomized experiments; at other times, practical circumstances, time pressures, and cost constraints necessitate compromises with the ideal. This chapter

EXHIBIT 7-A The General Problem of Assessing Program Effects

Estimating the effect of a new social or educational program requires comparing the condition of the individuals who have received the new service against the condition they would have been in had they not received the service. At times, their condition in the absence of the service is predictable; often, however, predicting how a group of individuals would have fared without the new service is difficult or impossible. A forecast of a group's behavior would, for example, have to take into account ordinary growth, cyclical or seasonal variations in behavior and the environment, and ordinary random fluctuations. Such a forecast also would need to determine whether the group might have received no services or services other than the new one, then somehow predict the effect of these unreceived services or alternative services.

In the absence of reliable predictions about a group's behavior, it is natural to construct a comparison group that has not received the new service. For a comparison to be fair, the comparison group must not differ systematically from new service recipients in any respect that would affect their future state. That is, the groups must be such that an unbiased estimate of the relative effect of the service is possible. More precisely, a fair comparison requires that the characteristics of individuals who receive services, or those who do not, be independent of the response variable used to make judgments about relative effectiveness. In other words, how people are selected for groups, or select themselves into groups, must not depend on factors that could influence outcome.

SOURCE: Quoted, with permission, from Robert F. Boruch, *Randomized Experiments for Planning and Evaluation: A Practical Guide* (Thousand Oaks, CA: Sage, 1997), pp. 1-2.

provides an overview of impact evaluations and the alternative ways of estimating the effects of social programs; the three following chapters go into detail about undertaking such evaluations.

Prerequisites for Assessing Impacts

In earlier chapters, we have outlined the two prerequisites for assessing the impact of an intervention. First, the program's objectives must be sufficiently well articulated to make it possible to specify credible measures of the expected outcomes or the evaluator must be able to establish such a set of measurable outcomes. Second, the intervention should be sufficiently well implemented that there is no question that its critical elements have been delivered to appropriate targets. It would be a waste of time, effort, and resources to attempt to estimate the impact of a program that lacks measurable outcomes or that has not been properly implemented. An important implication of this last consideration is that interventions should be evaluated for impact only when they have been in place long enough to have ironed out implementation problems.

We cannot overemphasize the technical and managerial difficulties involved in undertaking the more rigorous forms of impact evaluation. The targets of social programs are often persons and households who are difficult to reach or from whom it is hard to obtain outcome and follow-up data. In addition, the more credible impact designs are demanding in both their technical and practical dimensions. Finally, as we discuss in detail in Chapter 12, evaluation research has its political dimensions as well. The evaluator must constantly cultivate the cooperation of program staff and target participants to conduct impact assessment while contending with inherent pressures to produce timely and unambiguous findings.

Linking Interventions to Outcomes

The problem of establishing a program's impact is identical to the problem of establishing that the program is a cause of some specified effect. Hence, establishing impact essentially amounts to establishing causality. There are many deep and thorny issues surrounding the concept of causality that need not concern us here. Rather, we shall accept the view that the world is sufficiently orderly for research to yield valid statements such as "*A* is a cause of *B* in circumstances *C*." Note, however, that this statement recognizes that a given social phenomenon may have more than one cause, and usually does.

In the social sciences, causal relationships are ordinarily stated in terms of probabilities. Thus, the statement "*A* is a cause of *B*" usually means that if we introduce *A*, *B* is more likely to result than if we do not introduce *A*. This statement does not imply that *B always* results if *A* is introduced, nor does it mean that *B* occurs *only* if *A* is introduced. To illustrate, consider a program designed to reduce unemployment, such as job training. If successful, it will increase the probability that participating targets will subsequently be employed. However, the likelihood of finding a job is related to many factors other than amount of training, including conditions and processes that have nothing to do with training programs, such as the economic condition of the community. Thus, although the introduction of a voluntary employment training program for unskilled adults should raise the level of participants' technical skills and thereby increase the likelihood that they will find employment, no training program, no matter how well designed, will completely eradicate unemployment. Some target adults will simply refuse to take advantage of the opportunity offered, and some willing participants will be unable to benefit for a variety of reasons, not least the number of vacancies in the labor market.

By the same token, other factors besides the training program might be responsible for reducing the unemployment rate of the program targets. Economic conditions may take a strong turn for the better so that more jobs open up; employers may decide to take on new workers with limited skills, experience, or questionable work records, perhaps because they think they can pay them lower wages than more highly trained workers; new firms with strong needs for unskilled workers may start up; other programs may be initiated that provide hiring incentives.

Assessment of a program's real effects is complicated further by biases in the selection of participants. For programs in which participation is voluntary, there is always the possibility that those who choose to participate will be the ones most likely to improve whether or not they receive the services of the program. Men and women who enter employment training programs are often those persons who are most motivated to obtain employment and thus are more likely to reach that goal than less motivated targets whether or not they receive the training program. Similarly, students awarded scholarships may do better academically than other students, but they may well have been more likely to succeed even if they had not received the scholarships. Other factors favoring the selection of some targets into a program may not reflect motivation or ability so much as opportunity. For example, those living near a well-baby clinic may be more likely to use it, or persons with good literacy skills may be more easily reached by printed publicity. In short, the same factors that lead to self-selection by some participants into a program may also account for their subsequent improvement, a change that can easily be mistaken as an outcome of the program.

Still another confounding factor is that other social programs might be in effect at the same time as the one under examination. While a job training program is being implemented, for example, other initiatives may provide special incentives for employers to hire the unemployed, on-the-job training opportunities may become more available, or special "sheltered" jobs may be created to enable workers to gain experience while learning. Thus, the assessment of whether a specific intervention produces the desired effect is complicated by the many other factors besides the program itself that affect the condition in question.

The critical issue in impact evaluation, therefore, is whether a program produces desired effects over and above what would have occurred either without the intervention or, in some cases, with an alternative intervention.

"Perfect" Versus "Good Enough" Impact Assessments

In many circumstances, it is difficult or impossible to conduct impact evaluations using what is, in ideal terms, the best possible research design. In some instances, the available design options are so far from the ideal that the evaluator must question whether to undertake the assessment at all, especially if meaningful results are unlikely.

Unfortunately, evaluators are confronted all too frequently with situations where it is difficult to implement the "very best" impact evaluation design. First, as we explain later in this chapter, sometimes the designs that are best in technical terms cannot be applied because the intervention or target coverage does not lend itself to that sort of design. For example, the circumstances in which randomized experiments can be ethically and practicably carried out with human subjects are limited, and evaluators must often use less rigorous designs. Second, time and resource constraints always limit design options. Third, the justification for using the best design, which often is the most costly one, varies with the importance of the intervention being tested and the intended use of the results. Other things equal, an important program—one that is of interest because it attempts to remedy a very serious condition or employs a controversial intervention—should be evaluated more rigorously than other programs. At the other extreme, some trivial programs probably should not have impact assessments at all.

Our position is that evaluators must review the range of design options to determine the most appropriate one for a particular evaluation. The choice always involves trade-offs; there is no single, always-best design that can be used universally as the "gold standard." Rather, we advocate using what we call the "good enough" rule in formulating research designs. Stated simply, the good enough rule is that the evaluator should choose the best possible design from a methodological standpoint after having taken into account the potential importance of the results, the practicality and feasibility of each design, and the probability that the design chosen will produce useful and credible results.

The application of this rule is discussed in greater detail at several points in later chapters. For the purposes of presenting an overview of research designs in this chapter, we will discuss them as if there were no constraints on the choice of design. It should be borne in mind that this perspective generally needs to be modified in practice.

Gross Versus Net Outcomes

As we noted earlier, the starting point for impact assessment is the identification of one or more measurable outcomes that represent the objectives of the program. Thus, in studying a program designed to increase adult literacy, the objectives of the program may be stated as increasing reading-level scores on a standard reading skills test. The program may be considered successful if, after the program, the participants' scores are higher than what would be expected had they not received the program.

A critical distinction must be made at this point between *gross outcomes* and *net outcomes*, more aptly called *net effects*. The gross outcome consists of *all the change* in an outcome measure that is observed when assessing a program. Gross outcomes are usually easily measured and ordinarily consist of the differences between pre- and postprogram values on outcome measures. A gross outcome in an adult literacy program, for instance, would be any change in participants' reading level when measured just before participation in the program compared with measures afterward. In some cases in which preprogram values cannot be measured, gross outcomes may be measured in terms of postprogram values only.

Net effects are much more difficult to measure. Net effects are the changes on outcome measures that can be reasonably attributed to the intervention, free and clear of the influence of any other causal factors that may also influence outcomes. Gross outcomes, of course, include net effects but also include other effects that are not produced by the intervention. In symbolic terms, the relationship between gross outcomes and net effects can be expressed as follows:

$$\text{Gross outcome} = \left[\begin{array}{c}\text{Effects of}\\ \text{intervention}\\ \text{(net effect)}\end{array}\right] + \left[\begin{array}{c}\text{Effects of}\\ \text{other}\\ \text{processes}\\ \text{(extraneous}\\ \text{confounding}\\ \text{factors)}\end{array}\right] + \left[\begin{array}{c}\text{Design}\\ \text{effects}\end{array}\right]$$

Thus, a gain in literacy in before-and-after measurements of a group of participants in an adult literacy program (gross outcome) is composed of three parts: first, the effects of the program (net effect); second, the effects of extraneous confounding factors (consisting of selection effects and other events, experiences, or processes that influenced literacy during the period in question); and third, design effects (artifacts of the research process itself, including such factors as errors of measurement, sampling variations, and inconsistency in data collection).

Of course, impact assessments are concerned primarily with net effects. In the following two sections, we first discuss the problem of extraneous confounding factors and then turn attention to design effects.

EXTRANEOUS CONFOUNDING FACTORS

Given that gross effects reflect not only the effects of an intervention but also the effects of other processes occurring at the same time and already under way at the start of the intervention, the primary challenge of impact assessment is to arrive at an estimate of net intervention effects. To accomplish this, the evaluator must exclude or purge the confounding factors from the gross effects. That is, the influence of any extraneous factors that explain, in whole or in part, the observed changes in the target problem or population must somehow be removed from the estimates of the intervention effects. Note that these confounding factors are extraneous only in the sense that they are not wanted in the estimate of net effects. In other regards, they are often the factors that ordinarily produce the outcome, for example, natural recovery processes, changes in the economy, and the like.

Confounding factors vary according to the social phenomenon in question. An intervention designed to improve the nutritional habits of families will compete with processes quite different from those affecting a program to improve young people's occupational skills. Despite the idiosyncratic features of programs and their target populations, however, certain processes are general enough to be identified as potential competitors with almost any intervention. Some of the most important of these are outlined below.

Uncontrolled Selection

By *uncontrolled selection* we mean processes and events not under the researcher's control that lead some members of the target population to be more likely than others to participate in the program under evaluation. When there are preexisting differences between those who receive a program and otherwise eligible persons who do not, differences in the outcomes for these respective groups may be

accounted for by selection and not be attributable to the intervention. Such preexisting differences, when related to outcome variables, are known as *selection bias*.

Uncontrolled selection is among the most difficult of the extraneous confounding factors. Even if some person or agency deliberately chooses the targets for participation, such selection is still uncontrolled in the sense that the evaluator cannot account for it in a manner that allows its influence to be differentiated from true intervention effects. If the participants in a program are volunteers, uncontrolled selection is almost inevitable because those who volunteer almost certainly are more interested, more motivated, more appropriate or otherwise importantly different in relation to the program than those who do not volunteer.

Although the most familiar uncontrolled selection process is self-selection by targets who choose, of their own accord, to participate in a program, selection may come about in a variety of ways. As noted earlier, program location and access often play a role, as do motivation and such factors as whether prospective targets read newspapers and so learn about programs described there. In some "voluntary" programs, selection into a program may involve little choice from the viewpoint of participants as a result of political or administrative actions. Consider a community that through its municipal government, "volunteers" for a program to improve sewage disposal infrastructure. Although individual community members do not volunteer to participate, all persons living in the area are subject to the program and its potential benefits. A community in which officials are more likely to volunteer may be more progressive, more affluent, or otherwise different from other communities in ways that affect outcome measures. Similarly, when a new textbook is adopted for use in elementary schools, individual pupils ordinarily do not choose to use the book; the volunteering is done by the school system. Nevertheless, from the standpoint of impact assessment, this is a form of selection.

Similarly, "deselection" processes work in the opposite direction to bring about differential attrition in program participation. Seldom is the participation of everyone who begins a program carried through to the end. Drop-out rates vary from program to program but are almost always disturbingly large. The individuals who leave a program may well be different in significant and relevant ways from those who remain. For one thing, those who are clearly benefiting from the intervention are likely to remain, whereas those who find it unrewarding or difficult are more likely to drop out. The consequence of attrition often is that the participants who stay with a program are those who may have needed the program least and were most likely to have improved on their own.

At the beginning of this chapter, we discussed the experimental model that underlies impact evaluations. To have a true experiment, evaluators must control the assignment of targets to participant and comparison groups and, indeed, make that assignment on the basis of a random process. Random assignment under the researcher's control is not only the optimal way to equate experimental and comparison groups but may vitiate the need to adjust for extraneous confounding effects altogether.

Endogenous Change

Social programs operate in environments in which ordinary or "natural" sequences of events inevitably influence the outcomes of interest. Such naturally occurring effects are termed *endogenous changes*. For example, most persons who recover from acute illnesses do so naturally, because ordinary body defenses typically are sufficient to overcome such illnesses. Thus, medical experiments testing a treatment for some pathological condition—influenza, say—must distinguish the effects of the intervention from the changes that would have occurred without the treatment. Because almost all influenza sufferers recover from the illness, an effective treatment may be defined as one that accelerates the recovery that would have occurred anyway.

The situation is similar for social interventions. A program for training young people in particular occupational skills must contend with the fact that some people will obtain the same skills in ways that do not involve the program. Likewise, assessments of a program to reduce poverty must consider that some families and individuals will become better off economically without outside help.

Secular Drift

Relatively long-term trends in the community, region, or country, often termed *secular drift*, may produce changes in gross effects that enhance or mask the net effects of a program. In a period when a community's birth rate is declining, a program to reduce fertility may appear effective because of that downward trend. Similarly, a program to upgrade the quality of housing occupied by poor families may appear to be effective because of upward national trends in real income that enable everyone to put more resources into their housing, thereby producing gross effects that favor the program.

Secular trends can also mask the impact of programs by producing contrary effects that cancel out an intervention's positive net effects. Thus, an effective project to increase crop yields may appear to have no impact when gross effects are observed if unfavorable weather during the program period created poor growing conditions. Similarly, a program to provide employment opportunities to released prisoners may appear to have no effects if it coincides with a depressed period in the labor market.

Interfering Events

Like long-term secular trends, short-term events can produce changes that artificially enhance or mask net program effects. A power outage that disrupts communications and hampers the delivery of food supplements may interfere with a nutritional program. A natural disaster may make it appear that a program to increase community cooperation has been effective, when in reality it is the crisis situation that has brought community members together.

Maturational Trends

Evaluations of programs designed to produce changes in a target population must cope with the fact that maturational and developmental processes can produce considerable change that mimics or masks program effects. For example, the evaluation of an educational program designed to increase the language skills of small children must take into account the natural increases in such skills that occur with age. Similarly, the effectiveness of a campaign to increase interest in sports among young adults may be masked by a general decline in such interest that occurs when they enter the labor force. Maturational trends can affect older adults as well: A program to improve preventive health practices among adults

may seem ineffective because health generally declines with age.

Although several commentators have identified additional extraneous confounding factors (for classic discussions, see Campbell and Stanley, 1966; Cook and Campbell, 1979; Kish, 1987), these factors either are applicable primarily to laboratory conditions or are encountered rarely. The extraneous confounding factors we have listed are those to which evaluators must be particularly alert in designing impact assessment research. Strategies for isolating the effects of extraneous factors are discussed later in this chapter.

DESIGN EFFECTS

The obstacles to estimating net effects described so far are a consequence of the nature of the social problem involved, participant selection processes, and endogenous changes taking place in participants and in the program's social and historical context. These confounding factors are neither equally nor uniformly distributed across all impact evaluations. Thus, maturational effects may be of little concern in a study of the potential work disincentives of unemployment benefits, because the program is relatively short in duration and is directed at adults who are in the prime of their working lives. On the other hand, maturational effects are undoubtedly much more important in the study of the impact of long-term programs directed at preschool children or other age groups that naturally are changing rapidly.

Design effects, in contrast, result from the research process itself and thus are always present and always threaten the validity of impact assessments. Fortunately, our knowledge of many design effects is more complete than our understanding of extraneous confounding factors. Hence, whereas adjusting for extraneous confounding effects is always problematic, estimating and compensating for design effects are only sometimes problematic.

Stochastic Effects

The end product of an impact assessment is an estimate of net program effects based on empirical data. For example, a carefully controlled study of the impact of a teaching method may find that a class learning with that method increased its scores on an achievement test by 7.8 points more than a control group taught by conventional methods. The issue then becomes whether 7.8 points is a large enough difference to indicate a decided advantage to the tested teaching method, or whether a difference of this magnitude could be due to chance fluctuations.

Judgments about the sizes of differences between experimental and control groups are not easy to make, mainly because some differences can be expected independently of the effects of an intervention. An experimental class is likely to differ to some degree from its control even if both are taught in exactly the same way by the same teacher. Just as any random sample from a deck of cards in which there is exactly the same number of red and black cards will, through chance, often have an uneven number of reds and blacks, so too any two classes of students will differ from each other in learning when measured at any given point in time. Such chance-produced fluctuations, called *stochastic effects*, complicate the task of assessing the net effects of interventions.

Given the inherent instability of measures taken from samples, how can we judge whether a given difference is large enough that we would be safe believing it is not a chance fluctuation? Fortunately, mathematical models of sampling variation enable us to make that judgment rather precisely. Sampling variations are dependent mainly on two characteristics of the set of observations made: First, the larger the sample, the smaller the sample-to-sample variation; second, the more variable the individuals in a sample, the larger the sample-to-sample variation (in other words, the larger the standard deviation, or variance, the greater the sampling variability). Other factors, such as the expected shape of the distribution of sample-to-sample measures, also may play a role that is too complex to discuss in detail here. (Readers interested in pursuing this topic further should consult any of the many standard texts on statistical methods, such as Hays, 1990, or Myers and Well, 1995.)

For example, suppose that scores on an achievement test used to evaluate the effects of a teaching method had a normal distribution with an overall standard deviation of 10. If we use the test on two classes of 100 pupils each, sampling theory tells us that for two-thirds of classes of that size, sampling variations would lead to differences of between +0.7 and –0.7. Furthermore, only 1 comparison in 1,000 would show chance differences greater than +1.4 or less than –1.4. Given these considerations, it would be safe to assume that a difference of 7.8 score points between the two classes indicates an effect larger than can be reasonably attributed to the chance element in sampling variability. Such effects are said to be *statistically significant*, and this line of reasoning is known as *statistical inference*, that is, using what is known about the expected sizes of sampling variations to infer the likelihood that a given observation is due to chance.

Whereas chance fluctuations could lead us to conclude wrongly that an intervention had a certain effect, they may also make interventions appear ineffective when in fact they were effective. Thus, chance differences may make the impact seem larger or smaller than it actually is. The concept of *statistical power* is useful in understanding the issues involved here. Statistical power refers to the likelihood that a given impact evaluation design will detect a net effect of a given size, taking into account the statistical properties of the samples used and the statistical procedures employed to test for effects. For example, given an estimated value for the difference between posttest scores for experimental and control groups and the size of the sample, the probability of detecting the effect at a statistically significant level can be calculated. This is known as the power of the statistical analysis. Conversely, if an appraisal can be made of how large the net effects are expected to be, statistical power calculations can indicate how large the experimental and control groups must be for such effects to be detected reliably. These calculations are fairly straightforward; details and useful tables can be found in Cohen (1988), Kraemer and Thiemann (1987), and Lipsey (1990).

Using statistical inference to account for stochastic effects in impact evaluation involves making judgments about the relative importance of two types of error:

- Type I error (false positive): Making a positive decision when the correct decision is a negative one, that is, concluding that a program has an effect when it actually does not.
- Type II error (false negative): Making a negative decision when the correct decision is

positive, that is, failing to detect a real program effect.

The probability of making a Type I error is equal to the level of significance set for the statistical test used in the analysis. Thus, in testing a program that is in fact ineffective, one risks declaring it effective (false positive) 5 times in 100 trials if the significance level is set at .05. One can minimize false positives by setting a very strict criterion for statistical significance, but that increases the probability of making a Type II (false negative) error, for the two types of errors are inversely related for a given sample size. It is possible to minimize both types of errors simultaneously only by increasing the number of observations (sample size) or decreasing the variability in the observations through such techniques as statistically controlling the influence of covariates (Lipsey, 1990).

In every impact evaluation, the evaluator should decide a priori the importance of each of the two types of errors and then design the study and choose statistical analysis procedures accordingly. The judgment of whether it is more important to minimize false positives or false negatives should be based on substantive and practical grounds, not on theory or statistics. In testing the equipment of an airplane for safety, for example, it is clear that false positives are more serious than false negatives. That is, it is more important to avoid certifying as safe an airplane that might fail than it is to avoid rejecting as unsafe one that would not fail. One can make this judgment on the grounds that preserving life is more important than manufacturing airplanes inexpensively. Evaluating the safety of many medical interventions requires similar weighting of false positives.

In contrast, the opposite situation may apply in a relatively low-cost program, such as an educational television intervention. Effective educational programs of any type are difficult to design, and the negative effects of adopting an ineffective low-cost program are not very serious (especially in the absence of other educational alternatives known to be effective). It follows that false positives are less problematic than false negatives. It may be better to adopt a set of educational programs that, in statistical terms, are uncertain in their effectiveness in the hope that at least some actually are effective.

Tests of statistical significance are not the only basis for making judgments about the effects of an intervention. It is also useful to take into account evidence from other studies. Evidence from only one of several studies is less impressive than evidence that is consistent across studies. Indeed, replicating impact evaluations to determine whether the same effects are found is recommended when it is deemed important to be certain about effectiveness. In addition, it should be kept in mind that large samples can make small differences statistically significant when they are substantively unimportant and, conversely, small samples can make important effects statistically undetectable. These considerations have led to several suggestions to supplement statistical tests with other measures of intervention effects. For a discussion of these issues, see Browner and Newman (1989), Jacobson and Truax (1992), Kirk (1996).

Chapter 12 will take up the issue of judging evaluation results in greater detail. As we will see in that chapter, there are additional considerations, also based on value judgments, that should be taken into account. The important point here is that stochastic effects can be minimized to some extent by modifications in design. If evaluators anticipate small but important net effects, then sample sizes can and should be enlarged and variance control techniques applied so that the sampling variation will be smaller than the anticipated effects. If the new teaching method used as an illustration earlier was expected to produce average gains on an achievement test of between 0.5 and 1.0, and this range was meaningful in practical terms, statistical power should be high enough to ensure that they would be detected. This could be accomplished by increasing the numbers of students used in the design or by measuring and statistically controlling for variation from influential covariates, such as IQ, or both (see Lipsey, 1990). Because program stakeholders generally overestimate program effects, we recommend that considerable effort be made to ensure that impact designs have sufficient statistical power to detect statistically modest effects that may nonetheless be of substantive importance.

In this discussion, we have touched on only the barest essentials of the principles of statistical inference. Anyone planning to conduct impact evaluation should become familiar with the main issues and methods of statistical analysis, especially considerations of statistical power.

Measurement Reliability

A measure is reliable to the extent that, in a given situation, it produces the same results repeatedly. No measuring instrument, classification scheme, or counting procedure is perfectly reliable. Measurement error, or the extent to which a measuring instrument produces results that vary from administration to administration when applied to the same (or comparable) objects, plagues all measurement, whether of physical or social objects.

Although all measurement is subject to reliability problems, measures have such problems to varying degrees. Measurements of height and weight as obtained from standard devices and scales, for example, will be more consistent from one administration to another than measurements produced by repeated application of an intelligence test. By the same token, IQ tests have been found to be more reliable than reports of household expenditures for consumer goods, which, in turn, have been found to be more reliable than typical attitude scales.

For evaluators, a major source of unreliability lies in the nature of the measurement instruments used, many of which are based on participants' responses to written or oral questions posed to them by researchers. Differences in the testing or measuring situation, observer or interviewer differences in measure administration, and even participants' mood swings contribute to unreliability.

The effect of unreliability in measures is to dilute and obscure real differences. A truly effective intervention, the outcome of which is measured unreliably, generally will appear to be less effective than it actually is. An illustration of the effect of unreliability is shown in Exhibit 7-B. The table in that exhibit compares two measures of differing reliability in a hypothetical example of an educational intervention designed to raise levels of cognitive achievement among children from a disadvantaged background. The true outcome of the program is as shown in the top panel, labeled I. Forty out of 50 children (80%) in the participating group reached high achievement levels at the end of the program, but only 25 out of 50 (50%) of the nonparticipating or control individuals reached those levels. These are the results we would

EXHIBIT 7-B Hypothetical Example of Attenuation Effects of Measurement Unreliability on Intervention Outcomes

I. True outcome without measurement error

	Participants	*Nonparticipants*
High achiever	40 (80%)	25 (50%)
Low achiever	10 (20%)	25 (50%)
True program effect =	30%	

II. Comparison of results of measures of achievement that vary in reliability

	Correct Classifications			
	Measure A		*Measure B*	
	High	*Low*	*High*	*Low*
High achiever	60%	40%	90%	10%
Low achiever	40%	60%	10%	90%

III. Measured outcomes using Measure A and Measure B

	Measure A		*Measure B*	
	Participants	*Nonparticipants*	*Participants*	*Nonparticipants*
High achiever	28 (56%)	25 (50%)	37 (74%)	25 (50%)
Low achiever	22 (44%)	25 (50%)	13 (26%)	25 (50%)
Measured effect =	6%		24%	

observe if we had a perfectly reliable test of cognitive achievement, that is, a test that made no classification errors.

The reliability of two measures, A and B, is compared in the middle panel (II). Measure A is less reliable than Measure B. Note that whether a child is truly a high achiever or a low achiever, Measure A correctly classifies the child only 60% of the time. In contrast, Measure B correctly classifies 90% of the individuals measured. This means, also, that Measure A makes mistakes in classification 40% of the time, whereas Measure B makes such mistakes only 10% of the time.

The bottom panel (III) shows the different effects of the application of the two unreliable measures on the assessed outcome of the intervention. With Measure A, we find that, in total, 28 children (56%) are identified as high achievers. These 28 high achievers include 60% of the 40 children correctly classified as high achievers (i.e., 24 children) plus 40% of the 10 low achievers incorrectly classified as high achievers (or 4 children).

In contrast, with Measure B, 37 children are identified as high achievers (74%). They are composed of 90% of the 40 children correctly identified as high achievers (or 36 children) plus 10% of the 10 children (or 1 child) incorrectly so identified.

Using Measure A, we get a contrast between the nonparticipating and the participating groups of only 6%, whereas for Measure B the contrast is 24%. Obviously, the more reliable Measure B comes closer to showing the actual extent to which the program was effective (30%).

Note that neither measure provides an accurate estimate of the hypothetical true program effect; both underestimate the true effect. This problem is known as "attenuation due to unreliability" and is well documented (Muchinsky, 1996; Schmidt and Hunter, 1996). In most cases, it is impossible to eradicate unreliability completely, although it is possible to make adjustments in results that take it into account if the degree of unreliability is known. The point of the example in Exhibit 7-B is to emphasize the importance of care in both the construction and the application of measurement procedures.

There are no hard-and-fast rules about acceptable levels of reliability. Measures generally lose their utility, however, when their reproducibility falls below 75% to 80%—that is, when less than 75%-80% of objects measured on two occasions with the same instrument are given the same scores (see, e.g., Mehrens and Lehmann, 1991, and Suen, Ary, and Covalt, 1990, for ways to estimate reliability).

Measurement Validity

The issue of measurement validity is more difficult to deal with than the problem of reliability. A measure is valid to the extent that it measures what it is intended to measure. Although the concept of validity is easy to comprehend, it is difficult to test whether a particular instrument is valid because for many, if not most, social and behavioral variables, no agreed-on standard exists. For example, an attitude scale measuring "attachment to employment" ideally might require as a validity test some behavioral measure of the extent to which an individual remains employed when working and seeks employment when unemployed, a measure that clearly would involve long-term observations. To complicate the issue, employment and job seeking are affected by other variables besides attachment, including employers' decisions and workers' health. As a result, neither steady employment nor long-term unemployment always reflects degrees of attachment to employment. Any measure based on long-term observations would have to take involuntary employment changes into account. This adjusted measure would also have to be one on which most social scientists concerned with studying labor force attachment could agree.

Although in principle it may be possible to collect the behavioral data needed to provide a benchmark against which to validate a measure, to do so is ordinarily impractical in view of the time and costs involved. Furthermore, not all social scientists who are concerned with a topic would accept any proposed standard as appropriate. For the concept of attachment to employment, for example, some would perhaps argue that expressed willingness to work overtime was a more appropriate standard.

In practice, there are usually a number of ways a given characteristic might be measured; for example, many different questions could be asked that would be related, at least conceptually, to the idea of attachment to employment. If everyone could agree on the best method of measuring it, then potential measures could be compared with this best measure. In the absence of a best measure, the question of

whether a particular measure or set of measures is valid is usually a matter of case-by-case argument.

With outcome measures used for impact evaluation, validity turns out to depend very much on whether a measure is accepted as valid by the appropriate stakeholders, including members of the scientific community. Among social researchers, there is general agreement that one or more of the following criteria must be met for a measure to be considered valid:

1. *Consistency with usage.* A valid measure of a concept must be consistent with past work using that concept. Hence, if one develops a new scale to measure attachment to employment, it should not contradict the usual ways the concept has been used in previous studies.
2. *Consistency with alternative measures.* A valid measure must be consistent with alternative measures that have been used effectively by other evaluators. That is, it must produce roughly the same results as these established measures or have sound reasons for producing different ones.
3. *Internal consistency.* A valid measure must be internally consistent. That is, if several data items are used to measure a concept, the several measures should produce similar results, as if they were alternative measures of the same thing.
4. *Consequential predictability.* Some measures implicitly or explicitly entail predictions. For example, a measure of "propensity to move" implies by its name alone that it predicts whether or not a person or household will move. For such a measure to be judged valid, it should in fact predict moving behavior. Although not all measures have such clearly implied predictability, many do, and such measures ought to be tested for an adequate degree of predictability.

These criteria are clearly conservative in the sense that they stress the use of existing measures as reference points and discourage innovation in measurement. This conservative bent, however, is mitigated somewhat by the last criterion, consequential predictability. If a proposed new measure can be shown to be a better predictor than a previously accepted measure, then it may justifiably supplant the earlier one.

Clearly, a useful measure must be both valid and reliable; reliability alone is a necessary but not sufficient criterion for selecting measures. However, because a measure cannot be valid unless it is also reliable, assessment of reliability can be a first test of a measure's validity.

Choice of Outcome Measures

A critical measurement problem in evaluations is that of selecting the best measures for assessing outcomes (Rossi, 1997). We recommend that evaluators invest the necessary time and resources to develop and test appropriate outcome measures (Exhibit 7-C provides an instructive example). A poorly conceptualized outcome measure may not properly represent the goals and objectives of the program being evaluated, leading to questions about its validity. An unreliable outcome measure is likely to underestimate the effectiveness of a program and could lead to incorrect inferences about the program's impact. In short, an irrelevant or unreliable measure can completely undermine the worth of an impact assessment by producing misleading estimates. Only if outcome measures are valid and reliable can impact

EXHIBIT 7-C Reliability and Validity of Self-Report Measures With Homeless Mentally Ill Persons

Evaluations of programs for homeless mentally ill people typically rely heavily on self-report measures. But how reliable and valid are such measures, particularly with persons who have psychiatric problems? One group of evaluators built a measurement study into their evaluation of case management services for homeless mentally ill clients. They focused on self-report measures of psychiatric symptoms, substance abuse, and service utilization information.

Psychiatric symptoms. Self-report on the Brief Symptom Inventory (BSI) was the primary measure used in the evaluation to assess psychiatric symptoms. Internal consistency reliability was examined for five waves of data collection and showed generally high reliabilities (.76-.86) on the scales for anxiety, depression, hostility, and somatization but lower reliability for psychoticism (.65-.67). To obtain evidence for the validity of these scales, correlations were obtained between them and comparable scales from the Brief Psychiatric Rating Schedule (BPRS), rated for clients by master's-level psychologists and social workers. Across the five waves of data collection, these correlations showed modest agreement (.40-.60) for anxiety, depression, hostility, and somatization. However, there was little agreement regarding psychotic symptoms (−.01 to .22).

Substance abuse. The evaluation measure was clients' estimation of how much they needed treatment for alcohol and other substance abuse using scales from the Addiction Severity Index (ASI). For validation, interviewers rated the clients' need for alcohol and other substance abuse treatment on the same ASI scales. The correlations over the five waves of measurement showed moderate agreement, ranging from .44 to .66 for alcohol and .47 to .63 for drugs. Clients generally reported less need for service than the interviewers.

Program contact and service utilization. Clients reported how often they had contact with their assigned program and whether they had received any of 14 specific services. The validity of these reports was tested by comparing with case managers' reports at two of the waves of measurement. Agreement varied substantially with content area. The highest correlations (.40-.70) were found for contact with the program, supportive services, and specific resource areas (legal, housing, financial, employment, health care, medication). Agreement was considerably lower for mental health, substance abuse, and life skills training services. The majority of the disagreements involved a case manager reporting service and the client reporting none.

The evaluators concluded that the use of self-report measures with homeless mentally ill persons was justified but with caveats: Evaluators should not rely solely on self-report measures for assessing psychotic symptoms, nor for information concerning the utilization of mental health and substance abuse services, since clients provide significant underestimates in these areas.

SOURCE: Adapted from Robert J. Calsyn, Gary A. Morse, W. Dean Klinkenberg, and Michael L. Trusty, "Reliability and Validity of Self-Report Data of Homeless Mentally Ill Individuals," *Evaluation and Program Planning,* 1997, 20(1):47-54.

estimates be regarded as credible. Bausell (1992) identifies some useful resources for locating existing assessment instruments.

In addition to being reliable and directly enough related to the goals of the program to be valid, a good outcome measure is one that is feasible to employ, given the constraints of time and budget. Suppose, for example, that a family planning program, whose goal is to reduce average family size in the community, considers the following alternatives for measuring outcomes:

- Proportion of couples adopting effective contraceptive practices
- Average desired number of children
- Average number of children in completed families
- Attitudes toward large families

These four possibilities do not exhaust all the measures that can reasonably be viewed as relevant to the goal of reducing fertility. But even among the four, there are variations in terms of ease of measurement, cost of data collection, and probable validity. Thus, although a reduction in the average number of children in completed families (i.e., those past childbearing) may be the best expression of the eventual goal of a program to reduce fertility, the use of that measure to define the outcome implies a long-term evaluation of considerable complexity and cost. In contrast, it may be easy to measure attitudes toward large families, proceeding on the assumption that the impact of a fertility reduction program will be reflected in low approval of large families. However, given what is known about the often weak and erratic relationship between attitudes and behavior, a downward shift in the average desirability of large families is likely to be a remote measure of the program's goals. Because changes in attitude may occur without a corresponding shift in fertility practices, the "consequential predictability" of such a measure is not likely to be very high.

Of our four alternative ways of measuring the outcomes of a family planning program, shifts in contraceptive practices may, on balance, be the best choice. The relevant behavior can be studied over a relatively short period of time, there are ample precedents for adequate measurements in previous research, and shifts in contraceptive practices are known to be directly related to fertility. As a further illustration, Exhibit 7-D displays a variety of outcome measures that were established for a program designed to prevent adolescents from initiating the use of tobacco.

As illustrated above, often the most valid outcome measures either cannot be obtained directly at all or can be obtained only at prohibitive expense. Under such circumstances, indirect measures, generally referred to as *proxy measures*, must be substituted. A proxy measure is one that is used as a stand-in for an outcome that is not measured directly. The selection of a proxy measure is clearly a critical decision. Ideally, a proxy measure should be closely related to the direct measure of the program objectives but be much easier to obtain. In practice, it is often necessary to accept proxy measures that are less than ideal. Although there are no firm rules for selecting appropriate proxy measures, there are some guidelines.

First, for objectives that are measurable in principle but too costly to measure in practice, previous research may include studies that test the worth of alternatives. For example, suppose we wanted to assess whether the jobs obtained by persons completing training programs are better than those the trainees would have found otherwise as part of a broad evaluation of the quality of life of families participating in a comprehensive program that had a large number of intervention elements and outcomes. In principle, the quality of jobs could be measured by some weighted combination of earnings, wage rates, steadiness of employment, working conditions, and other measurable attributes. To do so might require surveying the former trainees and their comparison group peers and developing an extensive battery of survey items. Instead of this long and expensive procedure, the family member who was going to be interviewed anyway by telephone might be asked one or two items about the job satisfaction of targets of the training program. This procedure might be justified by previous research showing satisfaction to be highly correlated with the other attributes of "good jobs."

Second, objectives that are expected to be reached in the far future can be represented by proxy measures that reflect intermediate steps toward those goals. For example, although the objective of a program on family fertility is to reduce average family size, that goal can be measured definitively only after the women in those families have passed through their childbearing years. Proxy measures that center on the adoption of practices that will reduce completed fertility are reasonable surrogates—for example, adoption of contraceptive practices or changes in desired family size.

EXHIBIT 7-D Program Outcome Measures

A community intervention to prevent adolescent tobacco use in Oregon included youth anti-tobacco activities (e.g., poster and T-shirt giveaways) and family communication activities (e.g., pamphlets to parents). In the impact assessment the outcomes were measured in a variety of ways:

Outcomes for Youths

- Attitudes toward tobacco use
- Knowledge about tobacco
- Reports of conversations about tobacco with parents
- Rated intentions to smoke or chew tobacco
- Whether smoked or chewed tobacco in last month and, if so, how much

Outcomes for Parents

- Knowledge about tobacco
- Attitudes toward community prevention of tobacco use
- Attitudes toward tobacco use
- Intentions to talk to their children about not using tobacco
- Reports of talks with their children about not using tobacco

SOURCE: Adapted from A. Biglan, D. Ary, H. Yudelson, T. E. Duncan, D. Hood, L. James, V. Koehn, Z. Wright, C. Black, D. Levings, S. Smith, and E. Gaiser, "Experimental Evaluation of a Modular Approach to Mobilizing Antitobacco Influences of Peers and Parents," *American Journal of Community Psychology*, 1996, 24(3):311-339.

Third, when proxy measures, or any outcome measures of uncertain validity, must be used, it is wise to use several such measures when possible. Multiple measurement of important outcomes potentially provides for broader coverage of the concept and allows the strengths of one measure to compensate for the weaknesses of another. It may also be possible to statistically combine multiple measures into a single, more robust and valid composite measure that is better than any of the individual measures taken alone. In a program to reduce family fertility, for instance, changes in desired family size, adoption of contraceptive practices, and average desired number of children might all be measured and used in combination to assess outcome.

The Hawthorne Effect and Other Delivery System Contaminants

In a famous "before and after" study conducted in the 1930s, researchers attempted to determine the effects of varying illumination on the productivity of women assembling small electronic parts (Roethlisberger and Dickson, 1939). It was discovered that any change in the intensity of illumination, positive or negative, brought about a rise in worker productivity. The researchers interpreted this effect as an artifactual result of conducting the research and it has since been dubbed the *Hawthorne effect* after the site where the study was conducted. Ostensibly, the researchers were studying the effects of varying illumination levels, but during the research, there was continuous observation of work-group members by researchers stationed in the assembly room. Roethlisberger and Dickson reasoned that the workers took the fact that they were being given so much attention by the researchers as a sign that the firm was interested in their personal welfare. The workers' response was to develop a high level of work-group morale and increase their productivity. Thus, the measured gross effect was a combination of the effects of the intervention (varying illumination), the delivery of that intervention (apparent concern on the part of management and the presence of researchers in the workplace), and the constant observation. Because productivity continued to increase throughout the duration of the study even though the workplace illumination was first increased and later decreased, the researchers concluded that the workers' increased productivity could not be a response to variations in the levels of lighting but was due to the continuous presence of the researchers themselves.

The Hawthorne effect is not specific to any particular research design; it may be present in any study involving human subjects. For example, in medical experiments, especially those involving pharmacological treatments, the Hawthorne effect is known as the placebo effect. This is why the evaluation of the effectiveness of a new drug usually involves both a placebo control, consisting of a group of patients who are given essentially neutral medication (sugar pills), and a control given the standard pill commonly prescribed. The effectiveness of the new drug is measured by how much more relief is reported from it in comparison to that reported by those who received either the placebo or the standard pill. The placebo group is required for comparison because participants are often affected simply by the knowledge that they are receiving treatment, irrespective of the efficacy of the treatment itself. The placebo control enables the researchers to identify and allow for this artifact of the research process.

It is possible to exaggerate the importance of the Hawthorne effect. Reanalyses of the original study (e.g., Franke and Kaul, 1978) have cast doubt on whether the data actually demonstrate any Hawthorne effect at all, and it may, in fact, be less important than once thought. One competing explanation is that the Hawthorne research occurred during the Great Depression, a time of severe unemployment and layoffs. Workers may have perceived the research activity as evidence that they were not going to be fired and thus worked enthusiastically and unusually productively. The important point, however, is that an intervention consists not just of the intervention administered but of everything that is done to the targets involved. Evaluation researchers must allow that the act of research itself is an intervention.

Our discussion of the Hawthorne effect underscores the fact that an intervention is rarely delivered in a pure form; it can rarely be separated from its context. Thus, counseling therapy for juvenile delinquents involves not only the therapist but also other personnel (e.g., the intake clerks), a setting in which the therapy is conducted, the reactions of the juveniles' peers who know of the therapy, and so on. Every aspect of the intervention delivery system, including the physical plant, rules and regulations, and the labeling of targets, can affect the outcome of the intervention, so much so that monitoring of the delivery of interventions almost always is a necessary adjunct to impact assessments.

Missing Information

No data collection plan is ever fulfilled to perfection. For a variety of reasons, almost all data sets are "perforated"; that is, they have gaps in them consisting of entirely missing cases or ones for which some portion of the required measures do not exist. In studies that follow up participants after intervention, for instance, some respondents move away and cannot be located, others refuse to provide more information, others become too sick or disabled to participate, and some may die—all events that result in missing data for some individuals. Even for individuals who have consistently participated, some parts of the data are often missing: Interviewers forget to ask questions, or respondents inadvertently skip over items on questionnaires or refuse to answer questions they regard as intrusive, irrelevant, or ambiguous.

Were missing data randomly spread across observations, their main effect would be similar to that of unreliability, namely, to obscure differences. But ordinarily that is not the case; persons lost to a study through attrition are often different from those who remain in ways that are related to the intervention outcome. For example, in experiments on the effects of welfare payments, families in the control group, who receive no payments, will be more likely to drop out. Similarly, persons who refuse to answer questions are often different in outcome-relevant ways from those who answer. For example, high-income respondents are more likely than low-income ones to refuse to answer questions about income. As Exhibit 7-E describes, high school students whose parents do not consent to their participation in data collection are also different from those who do.

To reduce these biases, alternative survey items or unobtrusive measures may be used. Also, various analytical procedures are available to estimate the extent of missing data biases and to impute estimates for purposes of analysis (Foster and Bickman, 1996; Little and Rubin, 1987).

EXHIBIT 7-E Missing Data Bias Related to Parental Consent in School-Based Surveys

The evaluators of a major community health promotion initiative in the western United States were presented with an informative opportunity to investigate the differences in evaluation outcome data associated with different forms of parental consent for surveys of high school students. Seventeen schools in six California communities either received the program or were in the control condition. Even though all the students in each school were thus research subjects, outcome data could be collected only on those students whose parents gave consent for them to complete surveys. Most school districts require one of two forms of parental consent. Active parental consent requires the parents to sign a consent form that is returned to the school. Passive consent involves notifying the parents of the survey and asking them to return a signed form only if they do *not* want their child to participate.

The California educational code requires active parental consent prior to asking students sex-related questions. The outcome survey for the evaluation covered a broad range of health-related topics, including sexual activity and contraception. Parents were sent a consent form via first-class mail that explained the study and gave them three response options: (a) Sign and return the form indicating permission for their child to be given a complete survey (active consent); (b) not return the form, which would indicate willingness to have their child be given a version of the survey that excluded the sex-related questions (passive consent); or (c) sign and return the form declining to give permission for their child to participate in the survey at all (fewer than 2% elected this option). This provided the opportunity to compare the characteristics of student respondents available for research through active consent with those available through passive consent on all survey items except the sex-related ones.

In both grades surveyed (9th and 12th), students with active parental consent were significantly more likely to be white, female, have a grade point average of B or above, live in two-parent households, have college-educated parents, and be involved in extracurricular activities. With respect to health status and risk-taking behavior, a significantly smaller proportion of students with active parental consent reported their health as less than "very good" and fewer reported irregular seat belt use. Among 9th graders the prevalence of current cigarette smoking was significantly lower in the active consent group. These latter differences remained after controlling for demographic variables. In addition, students with active consent were significantly more likely to report having seen the health promotion information from the intervention.

Thus, data from adolescents whose parents gave active consent for their participation in research involving sensitive subjects were not representative of all those who were exposed to the intervention. Evaluation research that was restricted to collecting data under active consent would therefore lose, as missing data, responses from students with distinctive characteristics whose parents do not provide active consent but would accept passive consent.

SOURCE: Adapted from C. Anderman, A. Cheadle, S. Curry, P. Diehr, L. Shultz, and E. Wagner, "Selection Bias Related to Parental Consent in School-Based Survey Research," *Evaluation Review*, 1995, 19(6):663-674.

Sample Design Effects

Most evaluation research is carried out on samples of potential or actual targets and nonparticipant controls. Findings from such research can be generalized to other groups—for example, all targets—only if the samples are properly designed and the design is then carried out with fidelity. Designing samples is a technical task, and most evaluators faced with a sampling issue of any magnitude would be well advised to involve a sampling statistician for the purpose.

The goal of a sampling strategy is to select an unbiased sample of the universe of interest. The first task is to identify a relevant sensible universe, that is, a population that includes those units (persons, households, firms, etc.) that are actual or potential targets of the program in question. Thus, a program designed to provide benefits to young males between the ages of 16 and 20 needs to be tested on a sample of that group.

The second task is to design a means of selecting a sample from the identified universe in an unbiased fashion. An unbiased selection procedure is one that gives each unit in the universe a known, nonzero probability of being selected. In practice, this often means that every member of the universe has an equal chance of being selected. There are many ways of designing such a selection strategy; additional details can be found in standard textbooks on the sampling of human populations (e.g., Henry, 1990; Kish, 1995).

The final task is to implement a sample selection strategy with fidelity; that is, persons who are supposed to be selected for the sample should in fact be selected. Rarely, if ever, is a sample of noninstitutionalized persons carried out without some selected individuals being missed. Indeed, most survey researchers are pleased when they are able to obtain cooperation from 75% or more of a designated sample. The Current Population Survey, administered by the Bureau of the Census, is reputed to have the highest response rates of all continuing surveys. It routinely gets cooperation rates in the high 90s, but this record is very exceptional. Cooperation rates can be affected strongly by the effort put into achieving contact with designated participants and by ardent persuasion, but such efforts add to the research expenses (see Ribisl et al., 1996, for useful tips on minimizing attrition).

Minimizing Design Effects

As we have seen, design effects are aspects of research design whose influence ordinarily is to diminish the capability of a given study to discern net effects when they actually exist. Careful planning of evaluations is the best antidote to design effects. In some cases—for example, when evaluations are planned that involve developing new measures—pretesting may be advisable to ensure that any outcome measures are sufficiently reliable and valid to respond to intervention effects. Attention must also be given to selecting representative samples that are large enough to provide adequate statistical power, measuring those target characteristics that may be appropriate for statistical control during analysis, and minimizing missing data problems.

DESIGN STRATEGIES FOR ISOLATING THE EFFECTS OF EXTRANEOUS FACTORS

As we noted earlier, the task of impact assessment is to estimate the difference between two conditions: one in which the intervention is

present and one in which it is absent. The strategic issue, then, is how to isolate the effects of extraneous factors so that observed differences can safely be attributed to the intervention.

Ideally, the conditions being compared should be identical in all respects, save for the intervention. There are several alternative (but not mutually exclusive) approaches to approximating this ideal that vary in effectiveness. All involve establishing *control conditions*, groups of targets in circumstances such that they do not receive the intervention being assessed. The following common approaches to establishing control conditions are discussed in detail in this and the next three chapters:

- *Randomized controls:* Targets are randomly assigned to an experimental group, to which the intervention is administered, and a control group, from which the intervention is withheld. There are sometimes several experimental groups, each receiving a different intervention or variation of an intervention, and sometimes several control groups, each also receiving a different variant, for instance, no intervention, placebo intervention, and "treatment as usual."
- *Regression-discontinuity controls:* Targets are assigned to an intervention group or a control group on the basis of measured values on a precisely identified selection instrument. Because the basis for selection is explicitly known, its relationship to outcome measures can be statistically modeled and separated from any remaining differences between experimental and control groups.
- *Matched constructed controls:* Targets to whom the intervention is given are matched on selected characteristics with individuals who do not receive the intervention to construct an "equivalent" group, not selected randomly, that serves as a control.
- *Statistically equated controls:* Participant and nonparticipant targets, not randomly assigned, are compared with differences between them on selected characteristics adjusted by statistical means.
- *Reflexive controls:* Targets who receive the intervention are compared with themselves using measurement before and after the intervention.
- *Repeated measures reflexive controls:* Also called *panel studies*, this technique is a special case of reflexive controls in which the same targets are observed repeatedly over time both before and after the intervention.
- *Time-series reflexive controls:* This technique is a special case of reflexive controls in which rates of occurrence of some event or other such social indicators are compared at frequent time points before and after the intervention.
- *Generic controls:* Intervention effects among targets are compared with established norms about typical changes in the target population.

Full- Versus Partial-Coverage Programs

The most severe restriction on the choice of an impact assessment strategy is whether the intervention in question is delivered to all (or virtually all) members of a target population. For programs with total coverage, such as long-standing, ongoing, fully funded programs, it is usually impossible to identify anyone who is not receiving the intervention and who in essential ways is comparable to the individuals who are receiving it. In such circumstances, the main strategy available is to use reflexive controls, that is, some form of before-and-after comparison. In contrast, some interventions will not be delivered to all the potential target population. Programs may lack the resources to serve the entire target population or their activities may be restricted to certain jurisdictions or geographical areas. Also, new programs or those that are to be tested on a demonstration basis ordinarily have only partial coverage, at least during their early stages.

In all likelihood, no program has ever achieved total coverage of its intended target population. Even in the best of programs, some persons refuse to participate, others are not aware that they can participate, and still others are declared ineligible on technicalities. Nevertheless, many programs achieve nearly full coverage. The Social Security Administration's retirement payments, for example, reach most eligible retired people. As a rule of thumb, when programs reach as many as four of five eligible units (80% coverage), a program has "full coverage" for the purposes of the present discussion.

The smaller the proportion of the target population that is not reached, the greater the differences are likely to be between those individuals who are covered and those who are not. For all practical purposes, almost all children between the ages of 6 and 14 attend school; those who do not suffer from temporary or permanent disabilities, receive tutoring at home from parents or private tutors, or are members of migratory worker families who move constantly from work site to work site. Hence, children who at any point in time are not enrolled in school are likely to be so different from those who are enrolled that no amount of matching or use of statistical controls will produce comparability of the sort needed for meaningful comparisons. Similarly, a recent report on Head Start evaluation strategies recommended that no randomized experiments be done on that program because so large a proportion of its target population is now covered by the program (Collins Management Services, 1990).

Nonetheless, some aspects of the impact of full-coverage programs can be evaluated, especially if the programs are not uniform over time or over localities. These differences provide the evaluator with some limited opportunities to assess the effects of variations in the program. Thus, evaluators might not be able to assess what the net impact of elementary schooling is (as compared to no schooling at all), but they can assess the differential impact of various kinds of schools and of changes in schools over time. Because most educational policy issues revolve around improving the existing school system, impact assessments of proposed changes in that system may be exactly what is needed to inform policy decisions.

These variations in ongoing, established programs occur in a variety of ways. Policies change over time, along with their accompanying programs. Program administrators institute modifications to meet some new condition or to make administration easier. Thus, from time to time, Social Security benefits have been increased to take into account new conditions or to add new services (e.g., Medicare). Similarly, sufficient local autonomy may be given to states and local governments that a program (e.g., Temporary Assistance to Needy Families [TANF]) may vary from place to place. With proper precautions, such "natural variation" can provide a point of leverage for estimating some program effects.

For partial-coverage programs, a greater variety of strategies is available. If it is practical, as may be the case especially in new or prospective programs, the ideal solution is to use randomized controls. In this strategy, a set of potential targets, representative of those who might be served if the program goes full scale,

is selected by an unbiased procedure and randomly assigned to an experimental group and a control group. With sufficient numbers of persons, the process of randomization maximizes the probability that the groups are equivalent; that is, individual variations in extraneous variables are likely to be distributed across the groups in such a way that the experimental and control groups will not differ materially in ways related to the intervention outcome. When an evaluator cannot use randomization procedures in forming experimental and control groups or conditions, other types of control groups often may be formed from uncovered targets, provided that proper procedures are used.

A CATALOG OF IMPACT ASSESSMENT DESIGNS

The simultaneous consideration of control conditions, intervention features, and data collection strategies produces the schematic classification of impact assessment research designs shown in Exhibit 7-F. The designs are classified into those that are appropriate primarily for impact assessments of partial-coverage programs and those that are useful primarily for full-coverage programs. The following discussion examines each of the research designs shown in that exhibit.

Designs for Partial-Coverage Programs

Design IA: Randomized Experiments

The essential feature of true experiments is random assignment of targets to treated and untreated groups constituting, respectively, the experimental and control groups. In evaluation efforts, randomized experiments are applicable only to partial-coverage programs. Randomized experiments can vary greatly in complexity, as the following examples illustrate:

- Alarmed by a rapid rise in the number of children placed in foster care after being abused or neglected, the state of Illinois instituted a family preservation program consisting of intensive casework with families of abused and neglected children and contracted with the Chapin Hall Center for Children at the University of Chicago to evaluate its effects. Families at risk of having their children placed in foster care were randomly assigned to an experimental group who experienced the family preservation program or a control group who experienced "ordinary" child protective services, typically a much less intensive case work regimen. Both experimental and control families were tracked through repeated interviews and administrative records to ascertain subsequent foster care placement and abuse or neglect complaints (Schuerman, Rzepnicki, and Littell, 1994).

- To assess the impact of enriched preschool experience on school performance and adult functioning, in 1962 researchers randomly assigned low-socioeconomic-status three- and four-year-old children to an experimental group, who were enrolled in an intensive preschool enrichment program, and a control group, who were not enrolled. The members of both groups were studied throughout their schooling and into adulthood with the latest observations made when the participants were age 27. In their young adulthood, members of the experimental group were found to have higher incomes, more steady employment, and fewer arrests (Schweinhart and Weikart, 1998).

EXHIBIT 7-F A Typology of Research Designs for Impact Assessment

Research Design	*Intervention Assignment*	*Type of Controls Used*	*Data Collection Strategies*
I. Designs for partial-coverage programs			
A. Randomized or "true" experiments	Random assignment controlled by researcher	Experimental and control groups randomly selected	Minimum data needed are after-intervention measures; typically consist of before, during, and after measures
B. Quasi-experiments			
1. Regression-discontinuity	Nonrandom but fixed and known to researcher	Selected targets compared to unselected targets, holding selection constant	Typically consists of multiple before- and after-intervention outcome measures
2. Matched controls	Nonrandom and unknown	Intervention group matched with controls selected by researcher	Typically consists of before- and after-intervention measures
3. Statistically equated controls	Nonrandom and often nonuniform	Exposed and unexposed targets compared by means of statistical controls	Before-and-after or after-only intervention outcome measures and control variables
4. Generic controls	Nonrandom	Exposed target compared with outcome measures available on general population	After-intervention outcome measures on targets plus publicly available "norms" of outcome levels in general population
II. Designs for full-coverage programs[a]			
A. Simple before-and-after studies	Nonrandom and uniform	Targets measured before and after intervention	Outcome measured on exposed targets before and after intervention
B. Cross-sectional studies for nonuniform programs	Nonrandom and nonuniform	Targets differentially exposed to intervention compared with statistical controls	After-intervention outcome measures and control variables
C. Panel studies: Several repeated measures for nonuniform programs	Nonrandom and nonuniform	Targets measured before, during, and after intervention	Repeated measures of exposure to intervention and of outcome
D. Time series: Many repeated measures	Nonrandom and uniform	Large aggregates compared before and after intervention	Many repeated before- and after-intervention outcome measures on large aggregates

a. Many of these designs are also used for impact assessments of partial-coverage programs. This use is not recommended.

The critics of Aid to Families With Dependent Children (AFDC) have maintained that the incremental payments given to poor families for each child enticed them to have additional children. The state of New Jersey introduced a "family cap" modification to its AFDC rules in 1992, which prohibited increases in

payments for children born after enrollment. To test the effectiveness of this rule, a randomized experiment was started in which a control group of about 3,000 families was subject to the old AFDC rules that increased payments for additional children, and an experimental group of about 6,000 families were subject to the family-cap rules. Births and abortions occurring to both groups were followed using administrative records (Camasso, Harvey, and Jaganathan, 1996).

- The effects of an HIV prevention program in New York City were examined by randomly assigning 151 adolescents to seven sessions, three sessions, or no sessions of small group instruction and role-play. The intervention procedures involved learning cognitive-behavioral strategies, social skills, and HIV-related information. Over the subsequent three months, the evaluators tracked the number of unprotected risk acts and number of sexual partners for respondents in each group (Rotheram-Borus et al., 1998).

- To test the effectiveness of reemployment training and job search programs to help workers whose jobs had been eliminated in industrial restructuring, more than 2,000 displaced workers at several sites in Texas were randomly assigned to job search programs, to combined job search and retraining programs, or to control conditions. The workers were followed over a period of time to ascertain subsequent employment and earnings experiences (Bloom, 1990).

- The Big Brothers and Big Sisters program pairs adult volunteers with youths from single-parent households for purposes of forging a friendship through which the adult mentor can support and aid the youth. During 1991-1993, all youths who came to eight selected agencies were randomly assigned to receive a Big Brother or Big Sister mentor or go into a waiting list control group. Both groups were followed for the next 18 months and assessed with regard to use of alcohol and drugs, aggressive behavior, theft, property destruction, school grades, and school attendance (Grossman and Tierney, 1998).

The most elaborate field experiments to assess program effects are longitudinal studies consisting of a series of periodic observations of experimental and control groups extending, in some cases, over years. For example, the largest field experiments every conducted, the negative income tax studies in the late 1960s and early 1970s, all employed the same basic longitudinal design while varying in the kinds of interventions tested and the length of time over which they were given, ranging from three to ten years. One of these, the New Jersey Income Maintenance Experiment (Kershaw and Fair, 1976; Rossi and Lyall, 1976), was designed with eight experimental groups, each of which was offered a slightly different income maintenance plan, and one control group. Eligible families were randomly assigned to one of the nine groups. Each participating family was studied over a three-year period through monthly income-reporting requirements, quarterly and annual interviews, and special reviews of income tax returns. Of course, during the three-year period, the experimental group families were given cash benefits as part of the income maintenance intervention; in addition, both experimental and control families were paid fees for completing interviews.

It must be noted that large-scale field experiments generally involve testing prospective national policies and hence are concerned with generalizability to the nation as a whole. Small-scale field experiments, less concerned with national generalizability, are appropriate and, as some of the examples above demonstrate, have frequently been used to assess the effects of more localized interventions.

Most randomized experiments are designed with at least preintervention and postintervention measurement of outcome. The main reason for using both measures is to hold the starting points of targets constant in subsequent analyses of experimental effects. (There are also important statistical reasons for doing so, as is explained more fully in Chapter 8.) However, preintervention measures often are impossible to obtain. For example, prisoner rehabilitation experiments designed to affect recidivism can be based only on postintervention measures, because recidivism cannot be identified before release from prison. Similarly, intervention efforts designed to reduce the incidence of disease or accidents have undefined preintervention outcome measures. Several examples of post-only experiments are given in Chapter 8.

Design IB: Quasi-Experiments

A large class of impact assessment designs consists of nonrandomized quasi-experiments in which comparisons are made between targets who participate in a program and nonparticipants who are presumed similar to participants in critical ways. These techniques are called quasi-experimental because, although they use "experimental" and "control" groups, they lack the random assignment to conditions essential for true experiments. The following examples of quasi-experiments illustrate some of the nonrandomized controls to be discussed in this section:

- The Personal Responsibility and Work Opportunity Reconciliation Act of 1996 (PRWORA) profoundly changed public welfare, abolishing AFDC and substituting TANF. TANF is limited to five years of lifetime participation, emphasizes moving adult participants into employment, and is administered as block grants to states with wide discretion given to states to define their own programs. To monitor the effects of the program on poor families, the Urban Institute has instituted a telephone survey of some 50,000 households to be undertaken before and after the implementation of TANF that oversamples poor households and households with children. An additional survey is planned for 1999. Contrasting the findings of the two surveys provides a basis for assessing the effects of the changes from AFDC to TANF in the well-being of poor families and their children (Urban Institute, 1998).

- Births to poor women historically have been characterized by high incidences of neonate mortality, low birth weights, and high medical costs. The Special Supplemental Nutrition Program for Women, Infants, and Children (WIC), administered by the Department of Agriculture, provides supplemental food to pregnant women to help counter these adverse birth outcomes. To assess the effects of the program, more than 100,000 women who were Medicaid participants in five states during 1988 were studied. Using WIC and Medicaid records, the birth outcomes for women enrolled in WIC were compared with those who were not, statistically controlling for differences in age, education, marital status, and race of mother. Births to women enrolled in WIC were found to have a significantly higher average birth weight, lower mortality, and smaller Medicaid expenditures (Devaney, Bilheimer, and Schore, 1991).

• Using data on high school juniors and seniors gathered in a national sample of high schools in 1981, Coleman and his colleagues (Coleman, Hoffer, and Kilgore, 1981) found that students in Catholic high schools had higher achievement scores in mathematics and English than those in public schools. Using these data supplemented by a follow-up survey in 1983, and qualitative surveys conducted in a small sample of Catholic schools, Bryk and his colleagues were able to show that the advantages of Catholic high schools were due to the distinctive community climates of those schools and their uniform curricula. Furthermore, minority students enrolled in Catholic high schools did much better than their counterparts in the public schools (Bryk, Lee, and Holland, 1993).

• Success for All (SFA) is a program to improve instruction in the early grades so that all preschool and elementary students will have the skills necessary to succeed later in school. Its basic components include reading instruction, periodic assessments and regrouping for instruction, reading tutors, and family support. Students in one SFA school in Charleston, South Carolina were compared with those of another school chosen to be similar in student demographics and history of performance on district standardized tests. The scores on reading and math tests showed a positive effect for the SFA kindergarten program but inconsistent and small effects for the later grades (Jones, Gottfredson, and Gottfredson, 1997).

• A 15-week cognitive-behavioral skills training program for male spouse abusers was tested for effectiveness by comparing the 32 men who completed the program with the 36 who dropped out. Those who completed showed a lower rate of subsequent abuse than the controls (Hamberger and Hastings, 1988).

Four quasi-experimental designs are commonly used: regression-discontinuity designs, matched constructed control groups, statistically equated constructed controls, and designs using generic outcome measures as controls.

Regression-discontinuity designs. Evaluations that are based on regression-discontinuity designs come closest to the randomized experiment in ability to produce unbiased estimates of net intervention effects. Regression-discontinuity designs use a selection variable that must be strictly applied to determine placement in the intervention or control group. For instance, a *cutting point* may be defined on a pretest measure of reading ability to divide a sample into those with scores above that point and those with scores below. Those below (the poorest readers) are then given the reading program, and those above the cutting point are used as controls. The postintervention reading scores for the two groups are then compared while statistically controlling for the selection variable, leaving the groups otherwise comparable except for the intervention.

Regression-discontinuity analyses can be employed only for the assessment of programs in which the targets are selected, or can be selected, for intervention according to a sharp cutting point applied to scores on an explicit selection variable. For example, some college fellowship programs allocate awards on the basis of scores received on a standardized test (e.g., the National Merit Scholarship Competition). If the cut-off point is applied to those scores with reasonable consistency, good estimates of the net effects of receiving a fellowship can be derived by means of statistical analyses of differences in outcome measures around the cutting point, statistically adjusted for the relationship to the original selection variable.

Although this approach to studying impact is free of many of the problems associated with nonexperimental designs, its applicability is limited to those programs that can select participants on the basis of explicit, uniform, quantitative criteria. In addition, the statistical analysis required is sufficiently sophisticated that it cannot be used by persons without a relatively advanced knowledge of statistics. Chapter 9 discusses this design further; more discussion can be found in Trochim (1984) and Reichardt, Trochim, and Cappelleri (1995).

Matched "constructed" control groups. Historically, the "constructed" control approach has been the most frequently used quasi-experimental design. When two nonrandomly assigned groups are to be compared, it is generally better to use statistical controls to equate them rather than trying to match the groups case by case. However, the constructed control group might be used in circumstances where statistical control procedures cannot be undertaken because of untrained staff or unavailability of computer resources, or where insufficient data are available to support statistical controls. In the latter case, of course, the data deficiencies are also likely to make matching difficult.

Typically, in this design a group of targets is selected to receive an intervention, usually through normal program processes. To provide estimates of what their outcomes would be without intervention, the evaluator selects matching, unserved targets as controls who resemble the treated targets as much as possible in relevant ways. Relevant resemblance, in this case, refers to similarity on variables with important relationships to the selected outcome variables. The matched constructed control groups may be chosen from among existing groups, as when school classes are selected to match, in age and grade, a group of classes that are to receive a new educational program. Or they may be aggregates of individuals who are comparable to the targets receiving the intervention as when probationers receiving an intensive supervision program are matched with cases drawn from the files who did not receive that program.

Statistically equated constructed controls. A more sophisticated alternative to matching is provided by procedures that equate participants and nonparticipants by statistically "controlling" the role of variables on which they show initial differences in the analysis of outcome data. Typically, the equating is accomplished by using one of several multivariate statistical procedures, such as multiple regression, log-linear models, or analysis of covariance.

Typically, in this design a survey is undertaken of the target population or some sample of that population to identify targets who have and have not participated in a program and to obtain the data that will be used in statistically adjusting the two groups. To measure program impact, the researchers compare outcomes for participants and nonparticipants, statistically controlling for differences between the groups as identified by the control variables. Very sophisticated analyses of cross-sectional surveys, for instance, may attempt to model the processes by which participants are selected (or select themselves into programs, a topic dealt with in greater detail in Chapter 9).

Regression discontinuity, matched constructed control designs, and statistically equated constructed control designs are alternatives when the evaluator is unable to ran-

domize. Under favorable circumstances, they have the capability of removing the selection biases resulting from uncontrolled selection so that experimental and control groups can be meaningfully compared. In this regard, therefore, they resemble true experiments. They rely much more heavily than experiments on statistical models, and the assumptions required to apply those models, however, and thus are more vulnerable to error if those models or assumptions are not adequate.

Impact assessments using matching and statistical equating designs are also susceptible to whatever errors may be made in selecting the variables that are to be taken into account (i.e., adjusted for) in the comparisons between participants and nonparticipants. If important variables that differentiate the groups in relation to their likely status on outcome variables are not included in the statistical models, or are included but in distorted form because of poor measurement, the results may be biased. Thus, differences between the experimental and control groups on outcome might be due to inadequate statistical adjustments rather than to the effects of intervention.

Generic outcome measures as controls. Generic controls usually consist of measurements purporting to represent the typical performance of untreated targets or the population from which targets are drawn. Thus, in judging the performance of schoolchildren enrolled in a new learning program, the participants' scores on a standardized achievement test may be compared to published norms for schoolchildren of that age or grade. Although generic controls are widely available for certain subjects—IQ and academic achievement, for example—ordinarily they are not easily at hand. Furthermore, as discussed further in Chapter 9, generic controls are very often not suitable because targets are selected precisely because of the ways in which they differ from the general population on which the norms are based.

Designs for Full-Coverage Programs

Full-coverage programs present special difficulties to evaluators attempting impact assessments, because there are no unserved targets available to use as controls. As we discuss in more detail in Chapter 10, the only comparisons available to the researcher are between the same targets before and after exposure to the intervention, which are called reflexive controls, and between natural variations in such aspects as the activities, intensity, or duration of the program.

Although the designs discussed in the last section cannot be used for full-coverage programs, those discussed in this section could be employed to study programs with partial coverage. In particular, before-and-after designs without comparison or control groups are commonplace for partial-coverage programs. However, evaluators are strongly advised not to use them for that purpose. In most circumstances, the resulting impact estimates will not be credible because of the possibilities for bias resulting from various confounding effects such as maturation and secular drift.

Design IIA: Simple Before-and-After Studies

Although few designs have as much intuitive appeal as simple before-and-after studies, they are among the least valid of the impact assessment approaches. The essential feature of this design is a comparison of the same targets at two points in time, separated by a period of participation in a program. The differences between the two measurements are taken as an estimate of the net effects of the intervention. The main deficiency of such designs is that ordinarily they cannot disentangle the effects of extraneous factors from the effects of the intervention. Consequently, estimates of the intervention's net effects are dubious at best.

An additional complication is that when programs have been in place for a period of time, "before" measures normally can be gathered only by asking participating targets to reconstruct retrospectively what they were like before the intervention. In such studies, the unreliability of recall can be a serious design effect.

Design IIB: Cross-Sectional Studies for Nonuniform Programs

Although many full-coverage programs deliver a uniform intervention to all their targets, there are many in which the intervention varies. For example, all states have welfare programs, but the eligibility requirements and payment levels vary widely from state to state; indeed, the difference between payment levels in the least and most generous states is more than five magnitudes. The effects of these variations can be estimated using cross-sectional surveys that measure how much of an intervention is received (program dosage) and then contrasting measures of outcome for targets receiving different levels of intervention, perhaps with statistical controls for any important differences other than the program level.

Design IIC: Panel Studies (several repeated measures) for Nonuniform Programs

Panel studies are based on repeated measures of targets exposed to the intervention. Although panel studies appear to be a simple extension of before-and-after designs, the addition of more data collection points gives the results of these studies considerably more plausibility. The additional data at different time points, properly employed, allow the researcher to begin to specify the processes by which an intervention has impacts on targets.

This design is especially important in the study of full-coverage programs in which targets are differentially exposed to the intervention. In Chapter 10, we provide an example of how this design was used to study the impact of children's viewing of violence and aggression in television programs on their own aggressive behavior toward their classmates. Given the circumstance of almost universal television viewing among children and hence the virtual impossibility of establishing control groups who do not view TV, the best approach was to study how varying amounts of violent-TV viewing affected displays of aggression at subsequent points in time.

Design IID: Time-Series Analyses (many repeated measures)

Time-series consist of repeated measures taken on an *aggregate unit* with many data points preceding and following the point in time at which a new full-coverage intervention was introduced or an old program was substantially modified. By an aggregate unit, we mean periodic measures taken on a relatively large population or parallel samples of it, as, for

example, vital statistical series (births, deaths, migrations), crime rates, and economic indicators.

Although the technical procedures involved in time-series analysis are complicated, the ideas underlying them are quite simple. The researcher analyzes the trend before an intervention was enacted to obtain a projection of what would have happened without the intervention. This projection is then compared with the actual trend after the intervention. Statistical tests are used to determine whether or not the observed postintervention trend is sufficiently different from the projection to justify the conclusion that the intervention had an effect. For example, evaluators used a time-series analysis to study the effects of introducing community policing in Houston on calls for service, crime rates, and narcotics cases by analyzing the trends in these variables before community policing began and comparing them with the trends afterward (Kessler and Duncan, 1996).

Time-series analysis is especially important for estimating the net impacts of changes in full-coverage programs, particularly those that are delivered uniformly. In many full-coverage programs, every eligible target is given the same amount of the intervention. For example, most legislation, such as criminal codes, applies uniformly to all of its targets (i.e., all residents) in a given jurisdiction. Similarly, Social Security retirement payments are uniform for all persons with the same preretirement employment records. If retirement payments or sanctions for convicted felons are changed at some point in time, the impact of those changes can be studied through time-series analyses.

Time-series designs are the strongest way of examining full-coverage programs, provided that the requirements for their use are met. Some of the limitations of time-series analysis are detailed in Chapter 10. Perhaps the most serious limitation of time-series designs is the large number of preintervention observations needed to model preintervention trends accurately (more than 30 data points are recommended).

Indeed, a time-series analysis can be performed only if extensive before-enactment and after-enactment observations on outcome measures exist. Thus, it may be possible to study the effect of the enactment of a gun control law in a particular jurisdiction, but only if the evaluator has access to a sufficiently long-term series consisting of crime statistics that track trends in gun-related offenses over a long period of time. Of course, for many ongoing interventions such long-term measures do not exist. For example, there are no long-term, detailed time series on the incidence of certain acute diseases, making it difficult to assess the impact of Medicare or Medicaid on them. For this reason, time-series analyses are usually restricted to outcome concerns for which governmental or other groups routinely collect and publish statistics.

JUDGMENTAL APPROACHES TO IMPACT ASSESSMENT

Impact assessments using the designs outlined in Exhibit 7-F are often expensive and time-consuming. It is therefore tempting to turn to approaches that do not involve collecting new data or analyzing masses of existing data. In addition, circumstances may be such that none of the designs discussed can be used, especially when time pressures require net effect estimates within a month or two.

In this section, we discuss some of the major alternatives to the approaches presented so far. In these alternative approaches, the judgments of presumed experts, program administrators, or participants play the major roles in estimates of net impact.

Connoisseurial Impact Assessments

In connoisseurial impact assessments, an expert, or connoisseur, is employed to examine a program, usually through visits to the site of the program. The expert gathers data informally and renders a judgment. The judgment may be aided by the use of generic controls, that is, existing estimates of what the population as a whole usually experiences, or "shadow" controls, more or less educated guesses about what normal progress would be (see Chapter 10). Needless to say, connoisseurial assessments are among the shakiest of all impact assessment techniques.

Administrator Impact Assessments

Equally suspect are impact assessments that rely on the judgments of program administrators. Because of their obvious interest in making their efforts appear successful, such judgments are far from disinterested and impartial. Exhibit 7-G reports the findings of a team of evaluators who made an explicit comparison between the impressions of program staff about program impact and the results of an experimental impact assessment. Not surprisingly, staff viewed the program as having greater impact than the empirical evidence indicated.

Participants' Judgments

In the assessment of some programs, participants' judgments of program success have been used. These judgments have some validity, especially for programs in which increasing participant satisfaction is a stated goal. However, it is usually difficult, if not impossible, for participants to make judgments about net impact because they ordinarily lack appropriate knowledge for making such judgments.

The Use of Judgmental Assessments

Despite their obvious limitations, we do not mean to argue that judgmental assessments should never be used in estimating the impact of programs. In some circumstances, the evaluator can do nothing else. Although some might then advise against undertaking any assessment at all, we believe that some assessment is usually better than none. Evaluators may need to resort to judgmental designs when very limited funds are available, when no preintervention measures exist so that reflexive controls cannot be used, or when everyone is covered by a program and the program is uniform over places and time so that neither randomized nor constructed controls can be used.

QUANTITATIVE VERSUS QUALITATIVE DATA IN IMPACT ASSESSMENTS

Our discussion of research designs has so far been almost exclusively in terms of quantitative studies. Whether the data collected for an impact assessment should be qualitative or quantitative is a separate issue.

EXHIBIT 7-G Do Program Staff Have Exaggerated Impressions of Program Impact on Participants?

While evaluating drug education and prevention programs in junior high schools attended by students from high-risk neighborhoods, a team of evaluators interviewed program staff about their impressions of the programs' impacts on drug use behavior and risk factors related to drug use. On the drug use items, the overall staff response revealed some uncertainty about the impact of the programs—the majority indicated that they did not have a confident judgment about effects on youths' actual drug use. Still, rather substantial proportions said they believed that the programs had delayed clients' first use of drugs (25%-39%) and had generally prevented use of drugs by their clients (18-33%).

Staff views on the impact of the program on risk factors for drug use, however, were very positive. A majority (60%-90%) answered that the program had a distinct effect on participants' school attendance and performance, self-esteem, anger control, and peer and adult relations. Moreover, while most staff felt they had little opportunity to observe the youths' actual drug use, they did believe they were in a position to directly observe changes on these risk factors.

The impact assessment of the programs relied on self-report information gathered from program participants and control participants who attended comparable schools in the same communities. The evaluation results did not support the staff impressions about program effects. Responses from the participants showed that, relative to controls, there was little impact on their use of drugs, their attitudes toward use of drugs, or the various risk factors believed to be related to drug use.

Faced with this disparity, the evaluators considered two possibilities: Either there were program effects that the evaluation failed to detect but that were seen by staffers from their different vantage point, or there were in fact less substantial program effects than the staffers believed. Their conclusion was that the latter of these two possibilities was more likely. They found strong indications in their interviews that staff impressions were based mainly on anecdotal evidence of positive change in a few problem cases with which they were acquainted, change that may not even have been induced by the program.

Moreover, in this age group the actual rates of drug use and related problems are relatively low but staff nonetheless believed that the client population was at great risk for problematic behavior. The staff's faith in the efficacy of the program, therefore, might lead them to believe that these low rates were the result of the program's efforts. The evaluators noted that, in the interviews, staffers often asserted that without the ministrations of the program, many of their youthful clients would be using drugs.

SOURCE: Adapted from Steven A. Gilham, Wayne L. Lucas, and David Sivewright, "The Impact of Drug Education and Prevention Programs: Disparity Between Impressionistic and Empirical Assessments," *Evaluation Review,* 1997, 21(5):589-613.

Quantitative data are those observations that readily lend themselves to numerical representations: answers to structured questionnaires, pay records compiled by personnel offices, counts of speech interactions among coworkers, and the like. In contrast, qualitative data, such as protocols of unstructured interviews and notes from observations, tend to be less easily summarized in numerical form. Obviously, these distinctions are not hard and fast; the dividing line between the two types of data is fuzzy.

The relative advantages and disadvantages of the two types of data have been debated at length in the social science literature (Cook and Reichardt, 1979; Guba and Lincoln, 1994). Critics of quantitative data decry the dehumanizing tendencies of numerical representation, claiming that a better understanding of causal processes can be obtained from intimate acquaintance with people and their problems and the resulting qualitative observations (Guba and Lincoln, 1989; Lincoln and Guba, 1985; Patton, 1990). The advocates of quantitative data reply that qualitative data are expensive to gather on an extensive basis, are subject to misinterpretation, and usually contain information that is not uniformly collected across all cases and situations.

We cannot here resolve the debate surrounding data preferences. As we have indicated in previous chapters, qualitative observations have important roles to play in certain types of evaluative activities, particularly in the assessment of program theory and the monitoring of ongoing programs. However, it is true that qualitative procedures are difficult and expensive to use in many of the designs described in Exhibit 7-F. For example, it would be virtually impossible to meld a long-range randomized experiment with qualitative observations at any reasonable cost. Similarly, large-scale surveys or time series are not ordinarily built on qualitative observations.

In short, although in principle impact assessments of the structured variety shown in Exhibit 7-F could be conducted qualitatively, considerations of cost and human capital usually rule out such approaches. Furthermore, assessing impact in ways that are scientifically plausible and that yield relatively precise estimates of net effects requires data that are quantifiable and systematically and uniformly collected.

INFERENCE VALIDITY ISSUES IN IMPACT ASSESSMENT

The paramount purpose of an impact assessment is to arrive at valid inferences about whether a program has significant net effects of the desired sort. To accomplish this end, an impact assessment must have two characteristics: *reproducibility* and *generalizability.*

Reproducibility refers to the ability of a research design to produce findings that are robust enough that another researcher using the same design in the same setting would achieve substantially the same results. Generalizability refers to the applicability of the findings to similar situations that were not studied, for instance, similar programs in comparable settings.

Reproducibility

The reproducibility of an impact assessment is largely a function of the power of the research design, the fidelity with which the design was implemented, and the appropriateness of the statistical models used to analyze the resulting data. Impact assessments that use powerful research designs with large numbers of observations and that are analyzed correctly will tend to produce similar results whoever conducts the research. In this regard, randomized controlled experiments ordinarily can be

expected to have high reproducibility, whereas impact assessments conducted with cross-sectional surveys or using expert judgments and shadow controls (see Chapter 10) can be expected to have low reproducibility.

Generalizability

In evaluation research, generalizability is as important a characteristic as reproducibility. Indeed, one classic evaluation text (Cronbach, 1982) asserts that generalizability is at least as important as any other design feature in applied social research. For example, a well-conducted impact assessment that tests a demonstration program under conditions that would not be encountered in the program's actual operation may show that the program would be effective under those special conditions, but these findings may not be applicable to realistic program circumstances. In practice, the problem of generalizability is an especially critical one in the assessment of a prospective program, because such evaluations are usually conducted with a trial version of the program administered by the researchers.

The generalizability of an impact assessment is affected by a number of factors. To begin with, the sample of target units should be an unbiased sample of the targets that will be or actually are the clients of the enacted program. It would make little sense to test a new method of teaching mathematics on classes consisting of gifted children if the program is being designed for use in average classes. Obviously, a method that produces fine results with gifted children may not work as well with children of lesser ability. Similarly, a program to help the unemployed that is tested only on unemployed white-collar workers may yield findings that are not generalizable to other types of unemployed workers.

Assessments of ongoing programs may likewise be faulty if they are based on an inappropriate sample of the population of clients. The testing of gun control measures in a state such as Massachusetts, where gun ownership in the general population is quite low, may not generalize to states such as Texas or Arizona, where levels of gun ownership are very high.

Issues of generalizability also concern the variants of the programs being tested in an impact assessment. Assessments of a test program administered by highly dedicated and skillful researchers may not be generalizable to programs administered by government workers who do not have the same levels of commitment and skill. For example, a randomized experiment run by researchers to test the effectiveness of a prospective program providing limited unemployment benefit coverage to released prisoners produced results that were quite favorable to the prospective policy. Unfortunately, replications of the experiment in Georgia and Texas that used state agencies to administer the payment program produced results considerably at variance with the earlier, experimental findings (Rossi, Berk, and Lenihan, 1980). In short, for impact assessments to be generalizable, the interventions tested must be faithful reproductions of the programs as they actually are or will be implemented.

Other aspects of impact assessments also involve issues of generalizability. Often impact assessments are made in settings that may not closely resemble those that will characterize the enacted program. If an income maintenance program is evaluated on a sample of poor clients in an economically depressed community, it may produce effects that reflect the community setting as well as the program and hence lack generalizability to other types of communities. Likewise, the results of an impact assessment of an educational program may reflect the environment of the particular school used in the evaluation.

Whether or not an impact assessment will have high generalizability is always an issue in the assessment of prospective programs. The program when enacted may have only slight resemblance to the program that was tested, or the coverage of an enacted program may emphasize clients that are different from those used in the evaluation. Changes of this sort sometimes occur because in drawing up the appropriate legislation lawmakers may seek to find a program definition that will be supported by a variety of interests, and hence incorporate features that the evaluators did not test. The best an evaluator can do is to test a range of prospective programs and hope that the enacted program will fall within the range tested.

Some commentators on evaluation design issues have suggested that there is an inherent trade-off between reproducibility and generalizability, arguing that powerful, reproducible designs often cannot be conducted at reasonable cost on a large enough scale to meet high generalizability requirements. Thus, evaluation researchers may have to choose between reproducibility and generalizability. In such cases, reproducibility has been suggested as the more appropriate goal. Other evaluation experts (Cronbach, 1982) accept the trade-off but emphasize generalizability as the more important form of validity for evaluations. Cronbach asserts that less rigorous impact assessment designs of high generalizability are more relevant for policy purposes than very rigorous designs with low generalizability.

Our own inclination is to question whether the alleged trade-off is always a constraint in the design of impact assessments. We believe that the trade-off constraint will vary with the kind of program being tested. An evaluator must assess in each case how strong the trade-off constraint is and make decisions appropriately. For example, a program that has a very robust intervention (e.g., transfer payments) need not be nearly so concerned with the generalizability of the intervention as a program of human services whose interventions are tailored to individual clients, a variety of intervention that tends to be much less robust. Similarly, reproducibility goals may be judged as more important for interventions that are controversial or that may have undesirable side effects.

Perhaps the best strategy is to envisage the assessment of prospective programs as proceeding through several stages, with the early stages stressing reproducibility and the later ones stressing generalizability. This strategy in effect presumes that it is initially important to identify programs that work under at least some conditions. Having found such programs, it is then necessary to find out whether or not they will work under the conditions normally to be encountered under enactment.

Pooling Evaluations: Meta-Analysis

In some program areas, the existing evaluation literature is so extensive that it may be possible to examine reproducibility and generalizability empirically. To the extent that evaluations of very similarly configured programs yield convergent results, reproducibility is demonstrated. To the extent that similar program effects are found over a range of program variations, types of targets, settings, sites, and the like, generalizability is demonstrated (Cook, 1993).

A systematic approach to representing and analyzing evaluation findings across studies is meta-analysis. Meta-analysis involves coding the estimates of program effects and various descriptive information about the programs and methods involved in producing those effects for each of a number of comparable studies. This information is then compiled into a database that can be analyzed in various ways. Most relevant for the present discussion is analysis of the variation in program effects across different evaluations. Such analyses can not only show the degree of convergence or divergence of findings but can examine the relationships between observed program effects and the characteristics of the programs and methods involved in the evaluations (Lipsey and Wilson, 1996).

Meta-analysis has increasingly been used to summarize and analyze findings across large numbers of evaluation studies. Lipsey and Wilson (1993) reported on 300 meta-analyses of programs based on psychological, behavioral, or educational interventions. Major meta-analyses have been conducted in such program areas as marital and family therapy (Shadish, Ragsdale, et al., 1995), prevention in mental health (Durlak and Wells, 1997), Title I education (Borman and D'Agostino, 1996), juvenile delinquency (Lipsey, 1992), substance abuse prevention (Black, Tobler, and Sciacca, 1998), and scores of others.

CHOOSING THE RIGHT IMPACT ASSESSMENT STRATEGY

Our discussion of impact assessment designs has been built around the research model of randomized experiments and has offered that design as the most rigorous of all for yielding credible conclusions about the effects of an intervention. Nevertheless, the assessment approach to be chosen in a particular circumstance depends on a variety of contextual factors. For some types of programs, randomized experiments are simply inapplicable. In other circumstances, time, funds, and skills may preclude an experimental approach. With proper care, the other designs described in this chapter can be used effectively though with somewhat diminished confidence. In the next three chapters, we present examples of all these approaches, detailing their advantages and limitations.

SUMMARY

- Impact assessment is undertaken to determine whether a program has its intended effects. Such assessments may be made at any stage of program development, from preimplementation policy making through planning, design, and implementation.
- Underlying all impact assessment is the research model of the randomized experiment, the most convincing research design for establishing cause-and-effect relationships. The experimental model depends on a comparison of one or more experimental (intervention) groups with one or more control (nonintervention) groups. Although many impact assessments cannot make use of a strict experimental technique, all impact assessment designs compare intervention outcomes with some estimate of what has occurred or would occur in the absence of the intervention.
- A major task of impact assessment is to disentangle the net effects of a program from the gross effects observed. Various research designs permit researchers to estimate and sometimes counteract the influence of extraneous factors and design effects.
- Among the extraneous factors that can mask or enhance the apparent effects of a program are uncontrolled selection or attrition of participants and endogenous changes such as secular drift, interfering events, and maturational trends. To assess the true impact of programs, evaluators must be aware of these potential confounding factors and attempt to eliminate them or compensate for their influence.
- Aspects of research design that can obscure or enhance apparent net effects include stochastic effects, measurement reliability and validity, poor choice of outcome measure, the Hawthorne effect and other delivery system contaminants, missing data, and sampling bias. Careful planning of a research design can counteract the influence of most design effects.
- Depending on the nature of an impact assessment and the resources available, evaluators can call on a varied repertoire of design strategies to minimize the effects of extraneous factors. Different strategies are appropriate for partial- and full-coverage programs, because in full-coverage programs no untreated targets are available to use as controls.
- A number of design options are available for impact assessments of full- and partial-coverage programs, respectively, ranging from randomized experiments to time-series analysis. Although the various designs differ widely in their effectiveness, all can be used if proper precautions are taken.
- Judgmental approaches to assessment include connoiseurial assessments, administrator assessments, and judgments by program participants. Judgmental assessments are less preferable than more objective designs, but in some circumstances, they are the only impact evaluation options available.

- Impact assessments may make use of qualitative or quantitative data. Although qualitative data are important for certain evaluative purposes, precise assessments of impact generally require carefully collected quantitative data.
- Two key characteristics of the results of impact assessments are reproducibility and generalizability. In some situations, evaluators may need to decide which value to maximize in the research design. One approach for prospective programs is to emphasize reproducibility in the early stages of assessments and generalizability in the later stages.
- Meta-analysis enables researchers to pool the results of many impact assessments and analyze them to explore reproducibility and generalizability empirically. The findings of meta-analyses can be useful for investigating the variability of effects among programs in particular service area and summarizing the findings of large numbers of impact evaluations.