EVALUATION

METHODS FOR STUDYING PROGRAMS AND POLICIES

SECOND EDITION

Carol H. Weiss
Harvard University

Prentice Hall, Upper Saddle River, New Jersey 07458

8

DESIGN OF THE EVALUATION

> People who write about methodology often forget that it is a matter of strategy, not of morals. There are neither good nor bad methods but only methods that are more or less effective under particular circumstances in reaching objectives on the way to a distant goal.
>
> —George C. Homans (1949, p. 330)[1]

The evaluator's task is to create a study design that fits the questions to be asked. The design indicates which people or units will be studied and how they are to be selected, which kinds of comparisons will be drawn, and the timing of the investigation. This chapter sketches a number of design options. It gives an indication of the requirements for using each design and their advantages and disadvantages.

The designs I discuss should not necessarily be viewed as separate design packages, total entities among which the evaluator picks. Rather, they are bundles of techniques that can be put together in different combinations. The evaluator develops strategies for selecting participants, assigning them to groups, making comparisons, and timing the data collection. The designs in this chapter and the next provide a guide to available and well-tried evaluation practices, but the evaluator uses her judgment in tailoring them to the situation at hand.

The designs in this chapter have been used primarily in quantitative studies. They are geared to the collection of data that can be expressed quantitatively and compared over time, but there is no intrinsic reason why they could not be used when qualitative data are collected. For example, it is not common in qualitative evaluations to choose respondents by random sampling or to include a comparison group that does not receive the program. Yet on occasion there may be plausible reasons for doing so.

[1]From George C. Homans, "The Strategy of Industrial Sociology," *American Journal of Sociology, 54* (1949): 330–337. Copyright © 1949 The University of Chicago Press. Reprinted by permission of The University of Chicago Press.

Qualitative evaluators might think more explicitly about some of the design choices they make, just as quantitative evaluators might take into account features that are common in the qualitative world, such as an emphasis on the context within which the program functions and participants' perspectives on their experiences. This chapter can give ideas that make sense in qualitative as well as quantitative evaluations.

Designing the Process Evaluation

The experiment and most of the other designs discussed in this chapter are appropriate for outcome evaluations. For process evaluations, the modes of inquiry are frequently informal, and designs tend to be more casual. In fact, process evaluation is not very different from what is often called monitoring. One key difference is that monitoring is done primarily on behalf of funders and other high-level officials to hold the program to account. They want to know what is going on in the program for purposes of oversight. Evaluations of program process, in contrast, are often conducted for the benefit of the program. Process evaluations help the program understand what it has been doing and how, and lead to reflection on how it might improve its operations. Another difference is that process evaluations are generally more systematic than monitoring and rely more on data and less on intuitive judgments. But monitoring and process evaluation are similar kinds of inquiry.

A process evaluation has to be designed, whether it uses quantitative or qualitative methods of investigation. First, the evaluator has to decide which sites to study. If the project operates at a single site, that decision is made by default. But if there are several sites, or smaller sites embedded within larger ones like classrooms within schools, she has to choose how to allocate energies. She also has to decide which people to query. Such a decision can be made opportunistically, selecting informants informally as chances arise and as people pass her along to other people to talk to. Or she can use systematic sampling strategies to be sure that she covers principals, teachers, guidance counselors, and librarians and, within the teacher group, to represent teachers not only of the core subjects but also the arts, vocational subjects, special education, and bilingual classes. Another decision she has to make, consciously or by the way, is the time periods at which data will be collected—over how long a period, at what intervals, and how intensively.

The notion of *designing* a qualitative inquiry used to be heresy; it was taken for granted that such studies took shape in the field as the investigator figured out what was happening. In recent years, qualitative investigators—at least some qualitative investigators—have become more receptive to the notion of design. A big difference between the first and second editions of Yin's *Case Study Research: Design and Methods* (1984, 1994) is the addition of a 35-page chapter explicitly on design. Where many early books on qualitative approaches scanted the subject, several newer ones make design an integral feature (e.g., Maxwell, 1996; Miles & Huberman, 1994).

A major advantage of qualitative study of program process is the opportunity to find the unexpected. The evaluator can follow the trail wherever it leads. She also learns which aspects of the program matter to staff and to participants, what elements of the program are salient to whom at which times. For example, even if a health maintenance organization is giving important categories of preventive care,

clients may grumble over the lengthy spells of waiting time and the obstacles to seeing the same physician at successive visits.

Qualitative study of program process can take account of context. It can lead to awareness of how physicians' time is allocated and why scheduled health-care appointments are always late. It can look into the nature of the setting—physical facilities, neighborhood environment, and financial resources. Only constraints on time and access will limit the range of data that the qualitative evaluator assembles.

When should an evaluator conduct a process evaluation through qualitative methods? (a) When she knows little about the nature of the program and its activities. Qualitative methods allow her to poke around, listen to people, and become familiar with the program staff and participants. (b) When the program represents a marked departure from ordinary methods of service. A truly innovative program may set in motion a stream of developments that no one would have foreseen. (c) When the theories underlying the program—implicitly or explicitly—are dubious, disputable, or problematic. In such a case, the evaluator is uncertain of what to look at and measure in a process evaluation. A wider gauge inquiry is worthwhile.

But the evaluator may already be familiar with this kind of program and the elements that should be watched. She may also want to be sure that the study collects *comparable* data across sites, informants, and time periods. She may, for example, want to have consistent and reliable data on the frequency and intensity of program service and its fidelity to the program's original intent. In that case she can turn to techniques that yield *quantitative* measures of program process.

One procedure is to design forms that staff fill out on a regular basis. The forms can ask them to enter the nature of the problem the client presented, the type of service they provided, dates and times of service, and any special features of the contact. Such information will cumulate over time to provide a running picture of the program in action. Another way to collect data is through interviews with program recipients and/or program staff. The evaluator can develop a survey form that asks structured and unstructured questions about the operation of the program. All respondents can be asked the same questions so that data will be comparable across respondents, or an unstructured survey can be conducted, with a flavor of the ethnographic, to give people an opportunity to tell their story from their own perspective.

Whichever method, or combination of methods, is used for process evaluation, the evaluator has to decide the frequency with which information will be collected. These decisions will depend in part on the specific nature of program processes, such as how routinized they are, and the frequency with which their content and structure are apt to change. Information about characteristics of program participants is not likely to change very often, but information about the services they receive may call for more frequent recording. A plan should be made, too, to capture any big unexpected shift in the way the program operates.

Designing the Outcome Evaluation

The underlying logic of evaluation design for outcome studies is twofold: (a) to compare program participants before they receive the program with their situation afterward in order to see if they have made gains on key outcomes and (b) to compare program participants with an *equivalent* group of people who did not receive the program (a randomly assigned control group) in order to see whether participants' gains exceed those made by nonparticipants. In Figure 8-1, the difference between *b* and *a* represents the outcomes for participants (let's call the difference *y*). The difference between *d* and *c* represents the outcomes for the control group (let's call that difference *z*); *z* is the change that the group in the program would have experienced if they had not received the program. Then *y* minus *z* is the *net outcome* of the program. It is the additional advantage that the program produced over whatever gain would have happened anyway.[2]

FIGURE 8-1 DIAGRAM OF AN EXPERIMENT

		Before	After	
Random assignment	Program participants	a	b	$b - a = y$
	Control group	c	d	$d - c = z$

If *y* is greater than *z*, the program has had a positive net outcome.

The next chapter delves more deeply into experimental design and the randomization procedures required to implement it. Here I introduce it as a standard against which other designs are often judged. The experiment is very good at ruling out the possibility that something besides the program led to observed effects (e.g., participants were getting older and would have improved anyway, or conditions outside the program were responsible). Experimental design provides a level of confidence in the internal validity of results that other designs often aspire to.

Despite the acknowledged attractions of the experiment, it is often difficult to implement. As we will see in Chapter 9, it requires that the evaluator have a strong voice in the assignment of eligible people into the program and into the control group. That condition is often difficult to obtain. It also requires constant vigilance to be sure that the groups stay intact, continue to provide data, and do not become contaminated (e.g., by the controls' exposure to program materials). Donald Campbell made a great contribution to the discussion of design when he argued that what is essential is not one particular research design but the ability to rule out rival explanations (other than the program) for any changes in outcome that are observed (Campbell & Stanley, 1966; Cook & Campbell, 1979). He and his collaborators listed a dozen or so "threats to validity"—that is, conditions that could conceivably cause the observed changes—and he advocated designs that guarded against them.

In the absence of experimental design, conditions other than the program can be responsible for observed outcomes. Chief among them are the following:

Selection. Program recipients were different from the people with whom they are being compared from the beginning; therefore, differences at the end may be due

[2]Sampling from a population never produces two groups that are absolutely identical. Random fluctuations have to be taken into account. Therefore, the evaluator has to make inferences about whether the difference between *y* and *z* is greater than the likelihood of random error. She tests the difference to see whether it is statistically significant.

to the fact that different types of people were selected or selected themselves into the program.

Attrition. The loss of people from the program group was different from loss of people in the comparison group; therefore, even if the two groups were alike at the start, the surviving remnants may be different and outcomes may reflect these differences.

Outside effects. Participants in the program and comparison groups may have been exposed to different experiences and conditions (besides the program) while the program was in session; differential exposure to extraneous events may affect outcome measures.

Maturation. The sheer passage of time, and the processes involved in growing older, may account for observed outcomes. For example, as children age, their cognitive processes develop and mature, and they may learn more even without the help of the program.

Testing. Taking a test once may teach people to be better at test taking the next time around; outcomes may reflect this improved capacity. Similarly, responding to questionnaires and interviews in the beginning may alert people to the subject matter so that they become more conscious of the topics, and increased familiarity can affect outcomes.

Instrumentation. If there is a change in the instrument used to collect data from the preprogram to postprogram time, outcomes may reflect the change in data collection tool or technique. If raters change their standards for rating, this too represents a change in instrumentation.[3]

If there is no reason to suspect that selection or attrition or outside events or any of the other threats are getting in the way, then experimental design is a luxury. If one or another of the threats to valid conclusions *does* exist, then the design needs to guard against that particular condition. Of course, threats to valid conclusions can exist that the evaluator is not aware of, and she has to take scrupulous care to find out what they are. But the evaluator need not try to attain a perfect design for the edification of her research peers and the glorification of her research reputation; she is doing an evaluation to inform the policy community about the program and to help decision makers make wise decisions. Her task is to provide the best information possible. What the evaluator needs to be concerned about is countering *plausible* threats to valid conclusions.

Important Concepts

Validity

Here is a definition that is relevant: validity. Earlier we met the term in discussing measures and recognized that it had to do with the extent to which the indicator actually captured the concept that it aimed to measure. When we discuss validity of study conclusions, there is an analogous meaning. Validity has to do with the reality or truth of conclusions; valid findings describe the way things actually are. These days, when social scientists are conscious of the complexity of the world and the impossibility of

[3]Campbell and Stanley (1966) and Cook and Campbell (1979) use different words for two of the concepts. They call attrition "experimental mortality," which has too deadly a sound for me. They call outside events "history."

finding one reality or one truth, the definition of reality is usually understood as meaning approximate truth, or as close as we can come to describing reality.

To understand discussions of design, it helps to appreciate two kinds of validity (Campbell & Stanley, 1966). One is internal validity, which refers to the causal link between independent variables (which, for example, describe the participants or features of the service they receive) and dependent variables (particularly the outcomes of the program).[4] Internal validity is concerned with whether the program is the agent responsible for observed effects, rather than external conditions, artifacts of the methodology, or extraneous factors. It indicates whether the relationship between program inputs and observed outcomes is *causal*.

The other kind of validity is external validity, or generalizability. It is concerned with whether the findings of one evaluation can be generalized to apply to other programs of similar type. If this one AIDS prevention program is successful, can we generalize this to other similar kinds of AIDS prevention programs? How far can we generalize the findings? To which class of AIDS prevention programs will they apply?[5]

Research designs vary in the degree to which they achieve internal and external validity. A randomized experiment, in which units are assigned randomly to program and control groups, leads to high internal validity. When the units (say, work groups or individuals) were very much the same at the beginning (a condition that random assignment generally achieves), and had much the same experiences during the period of the program except for receipt of the program itself, any differences that are observed between the two groups at the end are fairly certainly due to the program.

However, experimental design may not be very good at attaining external validity. Because of the need for careful controls to maintain experimental conditions, the program group in the experiment tends to get hothouse treatment, artificially different from what people in other programs of the same type would be likely to get. So results may not generalize very well to other reading programs, drug crackdowns, or physical rehabilitation efforts (Cronbach & Associates, 1980).

There is often a tradeoff between internal and external validity. Designs that do well on one dimension don't always do well on the other. The evaluator has to know which matters more in the particular case, so that she can choose a research design with confidence that it is appropriate to the situation.

Unit of Analysis

The unit that the program aims to change (say, a school) is not necessarily the same unit that is selected into the program (e.g., a classroom). Either of these—or another unit such as a teacher—may be the unit of analysis, that is, the unit that the evaluation measures and enters into statistical analyses.

[4]An independent variable is one which is assumed to be a cause of other variables (dependent variables). In evaluation, independent variables are characteristic of participants, program process, and setting, which are assumed to affect the dependent variable, participant outcomes.

[5]Cook and Campbell list four kinds of validity in their compendium. The two that I do not discuss are statistical conclusion validity, which means that the relationship (or lack of relationship) found between the independent and dependent variables is real (and not an artifact, e.g., of too small a sample size or unreliable measures) and construct validity of cause and effect, which refers to the truth of generalizations about higher order abstractions from the variables used in the study (such as generalizing from a single variable used to measure motivation, to the relation between (presumably all) motivation and effects).

The *unit that receives a program* is usually a person, but it can be a group, an organization, or a community. That is, the program can aim to change *not* the person but the performance of a larger unit—such as an agency (e.g., increased referrals to other service-giving agencies) or a neighborhood (denser networks of friendship and mutual help). Even when such a program works through individual people, it is geared to reforming the behavior and sometimes the norms and beliefs of the larger entity. Measures of process and outcomes, and program theory, reflect this concentration on the supra-individual level.

A program to teach household budgeting to young couples provides services to the couple and is interested in the way the couple behaves. A program to reduce child abuse in at-risk families would sample families and seek to reach conclusions about family behavior. A services integration project attempts to induce social service agencies to coordinate the services they provide to children and families. The agency is the recipient of the program. Agencies are sampled to participate in the evaluation, and appropriate measures of outcomes would be the extent to which agencies make changes in the way they deliver services. As another example, consider a radio campaign to encourage charitable giving. Here the community reached by the radio station receives the program. While the program aims to stimulate individuals to give, it is delivered at the level of the community. Its effectiveness will be judged by how much the entire listening area gives. Therefore, one relevant outcome measure is the total contribution received; another might be the percentage of households that contribute. Similarly a change in highway signs and signals is meant to speed the flow of traffic along the whole highway system. The speed at which traffic moves, or the number of vehicles that exit the system in a given time period, is a possible indicator of effectiveness.

Sometimes the program *does* aim to change the individual, but it cannot reach individuals one by one. The people come in clusters, such as students in classrooms. The program has to be delivered to the whole intact classroom. This would be the case, for example, with a new curriculum or a special training program that gives teachers knowledge and skills that they are expected to implement in the classroom. The students are not independent; they are members of the class. Again, the unit receiving the program is the group—that is, the classroom.

The program can be a change in technology, such as introduction of a new computer networking facility, whose purpose is to change the functioning of the whole organization. The unit receiving the program is the organization, perhaps a state agency or a university. The program in this case is delivered to the organization.

The *unit of sampling* is the entity that is selected into the program. In community collaboratives that attempt to create neighborhood change, the unit selected into the program is the neighborhood. In Patterson et al.'s (1992) study of a nutrition education program carried out in supermarkets, supermarkets were selected. In programs conducted in schools, classrooms are usually the unit sampled. There may be two stages of sampling: First, schools can be sampled, followed by choice of classrooms within the selected schools.

The unit for sampling usually matches the unit to which the program is delivered. In two-stage and three-stage sampling, the last stage is generally the one that receives the program. For example, take the evaluation of a program of pull-out services for children with disabilities in regular classrooms. The first stage of sampling selects schools, the second stage selects classrooms, and the third stage selects individual students with disabilities who are going to receive the special program.

The *unit of analysis* is the unit that figures in data analysis. It is the unit that the measures measure, and in statistical analysis, it is the one that is entered into tables and equations.

Researchers used to worry about the appropriate unit of analysis. When analysis is carried out at a higher level—say, at the level of departments within a hospital—it is not possible to reach sound conclusions about individuals within those departments. That is the ecological fallacy, drawing conclusions at one level of analysis and trying to apply it to another level.

The ecological fallacy used to roil the research waters, but it is of modest concern these days. The reason is the advance in statistical methods. Techniques of multi-level analysis make it possible to analyze data at several levels simultaneously. The analyst need not be stuck with the unit of program delivery or the unit of sampling in analyzing the data. She can examine all levels at once.

A matter that requires continuing care is the unit of measurement. If the evaluator wants to be able to analyze the data at several levels, she has to be sure that the measures are attached to the several units. For example, if she wants to examine student absences at the student and classroom levels, she has to define, collect, and aggregate the data for the student (number of absences for each student) and the class (percentage of absences for the classroom). If she has absence data only for the classroom, she cannot infer it for the students. A 10% absence rate for the class doesn't mean that each student was absent 10% of the time. That is the ecological fallacy.

A community-building program that attempts to build a sense of connection among neighborhood residents and encourage participation in neighborhood affairs can be evaluated at the level of the individual—who gets involved, how much time they spend on neighborhood affairs, etc.—but some people believe that its target is really the neighborhood. If one takes the neighborhood as the unit of analysis, outcome measures have to be aggregated to the neighborhood level—for example, proportion of people who participate, number and density of networks of mutual help, percentage of residents who believe the neighborhood is a desirable place to live.

The evaluator needs to pay attention to the units that she is working with. There will be times when the choice of unit will make a large difference in her ability to draw responsible conclusions. An example is La Prelle, Bauman, and Koch's (1992) evaluation of a mass media campaign to deter initiation of cigarette smoking by adolescents. The study was conducted in six treatment communities and four comparison communities. The evaluators used careful design parameters and selected a sample of youth in each community. They then made statistical adjustments on 10 sociodemographic and personality variables that were correlated with smoking in order to compensate for community differences. Although the unit of sampling was the community, the individual was the unit of analysis, and there was so much variation among communities that even after statistical adjustment, they could not detect effects. Their conclusion was that they should have used a different design, perhaps including more communities and using the community as the

unit of analysis.[6] They caution that intercommunity variance is likely to be so high that evaluators will have difficulty drawing conclusions about the effectiveness of an intervention at the level of individual; community differences overwhelm individual differences (La Prelle et al., 1992).

The important thing is that the evaluator think through the phenomena that she wants to study. She will want to consider the level of the unit that is selected to participate in the program (and evaluation), the way the program is operated, whether individuals act independently or whether their actions are strongly constrained by some larger group of which they are members. She also has to attend to the level of the conclusions she wants to be able to provide. Then she develops a rationale for selection of an appropriate unit with which to work.

Designs

The following sections describe a number of designs that evaluators have used, ranging from the simple to the complex. Each of them is useful for some purposes, and each of them has limitations. Many of the strengths and limitations have to do with internal and external validity. You can think of these designs as an available set of options, but don't think of them as prefabricated options that can be taken off the shelf and unthinkingly applied to the situation at hand. Each design incorporates subcomponents that can be stitched together in a variety of ways (Cordray, 1986).

Study design takes place within the context developed in earlier chapters: key questions that have been posed, uses to which the study will be put, the program's theories of change, stakeholders' concerns. To create a design that is responsive to the needs of time and place, the evaluator's good judgment is the divining rod (Trochim, 1986).

As you read the following designs, you might keep in mind two questions: (a) What comparisons are being made, and will these comparisons provide sound conclusions? (b) Will the findings from a study like this be persuasive to potential audiences?

Informal Designs

Self-Evaluation

Perhaps the simplest way to evaluate a program is to ask the people who are involved with it what they think: staff, administrators, and clients. Staff are knowledgeable about what goes on and have day-to-day inside experience with activities and, often, with outcomes. They can render judgments about what is going more and less well and provide important suggestions for how to improve program activities.

Administrators, too, are insiders, and they have considerable knowledge about the program's pluses and minuses. In addition, they often get feedback (wanted and unwanted) from outside about what the press, the community, and other organizations like and dislike about the program.

Those are useful sources of information. When the program agency is collecting data to improve its own performance, such information may suffice. Its collection brings together the experience and practitioner wisdom of people who spend their working lives engaged with the program's vicissitudes, and their varied sources of knowledge can shed considerable light on the program.

[6]Their other suggestions were to use a stratified sampling technique, pairing communities at the outset, assigning one to treatment and one to control, or using time series designs.

But when the evaluation is destined for outside eyes as well, this kind of information is often suspect. Staff and administrators have a stake in the program. They may interpret subtle cues as progress where others would see little change at all. When staff members know that the evaluation is going to be reported to sponsors and funders of the program, they will generally seek to show the program in its most favorable light. Even if they don't *purposely* slant their replies to the evaluator's questions, they probably are more alert to the program's successes than they are to its shortcomings.

Clients of the program can also be asked to evaluate the program. They, too, are insiders and know a great deal about program workings. But they may judge the program on criteria different from those that animate the program. Students may judge a class by how entertaining it is rather than on how much they learn; delinquents in an institution may judge the program by how early it allows them to be released rather than on the extent to which it socializes them to lawful norms and behavior. Moreover, clients often have a stake in the program's continuation. Many times they don't want to lose its services, whatever the outcomes, because the program is the only available resource for them. So they may tailor their answers to what they think outsiders want to hear.

Despite the weaknesses, these kinds of judgments have value. Training programs for professionals often ask participants to fill out forms rating the whole program and each session, indicating the extent to which the sessions were interesting, informative, responsive to their needs, and so on. The intent is to use responses to improve the content of the training and the allotment of time to different topics. The participants are presumably the people best qualified to give this kind of guidance.

Evaluation of this type is, in essence, a popularity contest. But often the purpose of evaluation is to find out not only whether people *like* the program; it is also to find out if it is doing them any good. Self-evaluation alone is not likely to fill the bill. But judgments of this kind can be incorporated into designs that include other data as well.

Expert Judgment

A step beyond the judgments of people engaged with the program is the use of a knowledgeable outsider. An expert can be called and asked to examine the program and give his judgment. The judgment can be about any phase of the program, from its recruitment practices or financial accounting to the outcomes for its beneficiaries.

Many assessments are made in this way. When a federal monitor goes out to visit a program that his office funds, he is trying to evaluate the extent to which it is doing what it promised to do in its application for funding, the extent to which it is following rules and guidelines, and/or the extent to which it is producing the kinds of outcomes that society wants and expects. Some programs call in their own expert consultants to review their activities and suggest tactics for improving their work.

A variant of the expert is the connoisseur (Eisner, 1991). Here the analogy is to the art critic or the wine connoisseur. Again a person with wide experience and

refined standards of judgment makes the assessment. The connoisseur notices subtleties, experiences nuances, and recognizes import. Taking its warrant from the arts rather than the sciences, connoisseurship conveys its insights through metaphor, poetry, or narrative.

School and college accreditation is done though the use of expert judgments. In this case, not one individual but a team of individuals visits an institution, inspects its records and physical plant, reviews the self-evaluation done by its faculty, and interviews administrators, faculty, and students. The team then makes a judgment about whether to accredit the school or program and whether to affix conditions to the accreditation. Even when accreditation is assured, the team will often make recommendations that the institution take steps to meet accepted standards.

Another area where expert judgments are commonplace is in the evaluation of research and development (R&D) programs. Individual research proposals are commonly judged by experts through a mechanism known as peer review, in which a panel of researchers in the field covered by the proposal reviews its quality and importance. The evaluation of R&D *programs* (i.e., the whole set of studies funded and/or conducted by a unit) poses extraordinary problems. Research studies by their nature are ventures into the unknown, and some proportion of them are bound to come up dry. Even when studies lead to new findings or theory, the value of the results will not be known for a long time. Some will add knowledge or change thinking in the field, but only a subset will lead to measurably valuable inventions.

With an intangible product, vague goals, and long time frame, the worth of R&D programs is hard to assess (Bozeman & Melkers, 1993). The best resource is often the opinions of knowledgeable people who have little or no stake in the program under review.

The basis for experts' judgments is their own prior experience. They have seen many other programs and projects, and in their minds they compare this program or institution to others that they know. They usually have some ideal standards as well, some ideas about what a good program should look like. Therefore, they can base their assessment on both comparative information and normative criteria.

The advantages of a team of experts over a single individual are that no one individual's idiosyncrasies carry the day; no single person can exercise his unique standards of judgment. Decisions have to survive the vetting of the team. Also, the team can call on a wider array of experience and skill. Persons who know a great deal about schools' science curriculums are joined by those who know about physical facilities and library practices.

Experts are probably better at judging the procedures and practices of a program than they are at judging outcomes. In assessing program *process*, they can draw on their wide knowledge of other programs and experiences. But they may have difficulty in judging whether the recipients of program service are doing better or more poorly than people in other programs or even people who get no program at all. Without good data on program outcomes, they have to make arbitrary judgments. If they assume that certain kinds of program activities are necessarily linked to good outcomes, the accuracy of their judgments of outcome hinges on whether the hypothesized link holds true. If they talk to a few clients, they may be drawing conclusions from an unrepresentative subset of the program population.

How convincing an evaluation by expert judgment will be to others depends in large part on the reputation of the expert. A widely acknowledged expert on alcoholism treatment might be widely believed. But almost every field has conflicting schools of thought, and one side's expert may be unacceptable to another side. While he may have no bias for or against the particular program under review, he may well have an ideological preference for a particular mode of program practice, and his judgments about the program will reflect his preference. Even the best qualified expert is using subjective standards, standards that he usually finds hard to articulate and that are open to dispute. If his report fails to satisfy some constituencies, they are apt to question his credentials, his experience, and even his integrity, as the testimony of experts in legal trials demonstrates. Sometimes a respected expert ventures into a program whose cultural and historical characteristics are beyond his experience, as, say, in rural Alaska, and his assumptions fail to take the setting into account. The expert's credibility usually depends on who he is rather than what he knows about the particular program under review. Most evaluations try to use methods that are less vulnerable to the vagaries of individual judgment.

Formal Designs

One-Group Designs

A commonly used set of designs examines a single program. The program may be an individual project or a set of projects or a whole statewide or federal program. With this design, the evaluation does not include any comparison with units (people, communities) that did not receive the program. Evaluations of this sort fall into two categories: those that look at units receiving the program only *after* the program has been in operation for a while or is completed, and those that look at it both *before and after*. The kind of data collected can be qualitative or quantitative.

After-only designs are usually employed when the evaluator isn't called in until the program is well under way and no before data are available. Before-and-after studies require an evaluator to be on the scene at the start. Whether data are collected on the situation prior to the beginning of the program or not, data can be—and often should be—collected on program process during the course of the program.

After Only Sometimes an evaluator isn't called in until the program is in midstream and there is no possibility of finding out what recipients looked like before. In such a case, the evaluator finds out what the situation is after the program and makes a series of assumptions about what things looked like before. On what can she base these assumptions?

First, she can examine records. For example, schools keep records on students that include much data on their backgrounds, prior school experience, grades, test results, health records, and so on. Hospitals, employment agencies, housing projects, and law enforcement agencies also keep files on clients. If the evaluator is lucky, the files include measures on the variables of interest for those who have been in the program at a prior point in time.

If records are unavailable, the evaluator can use historical comparisons. For

example, in evaluating a new history curriculum for ninth graders, the evaluator can compare test results at the end of the semester with history test results of the previous year's ninth graders—or (if last year's class was a bit quirky, and any single class is a bit quirky) the five-year average of ninth graders in the school. This adds useful information. But there are drawbacks if the situation in the school has changed. The school may be drawing its student body from a different population, the tests may be different, the teachers may have changed.

The evaluator can ask people involved with the program what the before situation was. Participants can be expected to recall their alcohol intake, job skills, or parenting practices at the time they entered the program. Relying on such retrospective reports seems reasonable enough, but it doesn't always produce valid data. People's memories are surprisingly unreliable. Usually people will remember the past as closer to the present than it really was or be overly influenced by one or two out-of-the-ordinary cases. Without records, they have real difficulty evoking a reliable baseline. But on certain matters of fact, such as age, number of years of schooling completed, and whether they were employed or unemployed and what kind of work they were doing, responses are fairly trustworthy.

Another way to elicit an estimate of the before situation is through the use of experienced judgment. Estimates can be made by well-informed people, perhaps the evaluator herself, about what participants were like at the outset, based on their own experience or prior research. The evaluator may take account of other evaluations she has done or refer to studies on groups similar to current program clients. Outsiders are likely to question the accuracy of such estimates (as would you and I). Without records on very similar populations, the imputation of preprogram data is shaky.

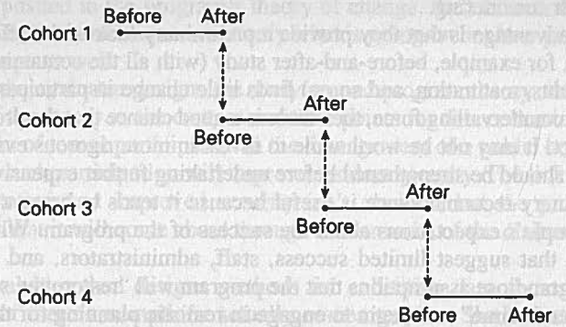
If at the time of the evaluation there are people in various stages of the program—some nearing the end, some part-way through, some just entering, data on the newcomers can be used to simulate a before measure. Provided that recruitment procedures have not changed and that there has not been a high rate of drop out during the program, the status of new entrants may be a reasonable basis for inference about the "graduates"—at least if numbers are large enough so that the laws of probability kick in.

Marsh and Wirick (1991) used this kind of design in their evaluation of a teenage pregnancy and parenting program at Hull House. See Figure 8-2. Each year a new cohort of young women entered the program at two sites in Chicago, who were described as a "new but basically similar group of clients." The post measures of group 1 were compared to pre measures of group 2, and so on for four years.

In this iterative fashion, if methods of recruiting participants have not changed, information is available to approximate the before-and-after design. Outside events are likely to be different for successive groups, and if the evaluator suspects that such events will influence the incidence of pregnancies, she can add a substudy to investigate this one issue. Similarly, additional measurements can be patched on to test whatever other rival hypotheses challenge the validity of evaluative conclusions.

Before and After The logic of this design seems clear cut: Look at program recipients before they enter the program and then again at the end. The difference between their condition at Time One (T1) and Time Two (T2) should be the effect

FIGURE 8-2 COMPARISON-GROUP DESIGN WITH SUCCESSIVE COHORTS



Note: Dotted lines indicate comparisons made between the after data of one cohort and the before data of the succeeding cohort.
Source: Adapted from Marsh & Wirick, 1991.

of the program. Oh, but is it? Only a moment's thought will reveal that many other things happen to program recipients besides their participation in the program. They watch television, get sick, fight with their parents, make new friends, move to a new house, take an evening school course. Was the change in their skills or health or income due to the program? Maybe, but then again, maybe not.

At their best, before-and-after studies can be full of detail, provocative, and rich in insight. If the data are collected with care and system, they offer considerable information about the program and the people who participate in it. When before and after data are supplemented by "during-during-during" data, the evaluation will be able to say much about how the program is working. If the evaluation systematically tracks the assumptions of program theory, it will be able to say much more.

For formative purposes (i.e., for modifying and improving the program), the design may be sufficient. But for summative purposes (i.e., for rendering judgment on how effective the program has been), it is not as authoritative as more rigorous designs. It provides no answer to objections that maturation or outside events were responsible for whatever change occurred. When before-and-after study shows no change, it is possible that outside factors tamped down real changes that would otherwise have been observed. For example, suppose there had been a house-building program for Little Pigs to teach them all how to build brick houses, but after measures taken some time after the conclusion of the program showed that no Little Pigs' houses were still standing. It may be that the program hadn't effectively taught brick-building skills, but it is also possible that the Big Bad Wolf had bought a pile driver that allowed him to level brick houses.

Advantages and Disadvantages of One-Group Designs As a general rule, one-group designs, while generating important information, leave room for differing interpretations of how much change has occurred and how much of the observed change was due to the operation of the program. Critics can launch effective attacks on the methodology and claim that results are indeterminate. They can point out the

threats to validity of conclusions. Still, with all the caveats, there are times when they are worth considering.

A first advantage is that they provide a preliminary look at the effectiveness of a program. If, for example, before-and-after study (with all the contaminating effects of outside events, maturation, and so on) finds little change in participants, and there is no obvious countervailing force, then there is a good chance that the program is having little effect. It may not be worthwhile to invest in more rigorous evaluation now. The program should be strengthened before undertaking further expensive inquiry.

Preliminary reconnaissance is useful because it tends to bring a modicum of realism to people's expectations about the success of the program. When confronted with data that suggest limited success, staff, administrators, and funders may reduce their grandiose assumptions that the program will "restore the sense of community" or "end crime" and begin to engage in realistic planning for the future.

Note that program effects may be *underestimated* if outside events operate to counteract program efforts, or if the evaluator in her mind's eye is comparing participants to groups who can be expected to do better—say, because of higher socioeconomic status or higher ability. If one-group studies do reveal change, and there is serious interest in the extent to which change is attributable to the program, further evaluation can be done under more controlled conditions.

A second reason for considering one-group designs arises from practices of agencies that fund evaluations of social programs. Some agencies demand one-time ex post facto investigation; they are responding to political pressures and short-term needs, and they want quick results. Evaluators will have to exploit every opportunity to supplement and expand the basically inadequate design.

Extending One-Group Designs

One-group designs can be elaborated in two basic directions: collecting more data on what happens during the program and collecting more data on what happens much before and much after the program.

More Data during the Program One-group evaluation need not be limited only to pretest and posttest measures. There are several ways to extend their reach even without adding a comparison group. One way is to take a "during" measure or a series of "during-during-during" measures. Quantitative or qualitative data can be collected not only on program services but also on participants' progress as they move through the program. These data can be analyzed, either quantitatively or qualitatively, to elaborate the picture of what happens during the program and to identify the association between program events and participant outcomes.

When the sponsor of the evaluation is interested only in the single project under study and not in generalizing to other projects, one-group evaluation can focus on the unique services, events, and people on the scene. It can probe into the relationships among rules, recruitment strategies, modes of organization, auspices, services, leadership, communication, feedback loops, and whatever else people are concerned about. Note that here we are moving in the direction of qualitative investigation.

A further move is to use program theory as the basis for data collection and analysis. Data on participants' progress are laid alongside the assumptions that the-

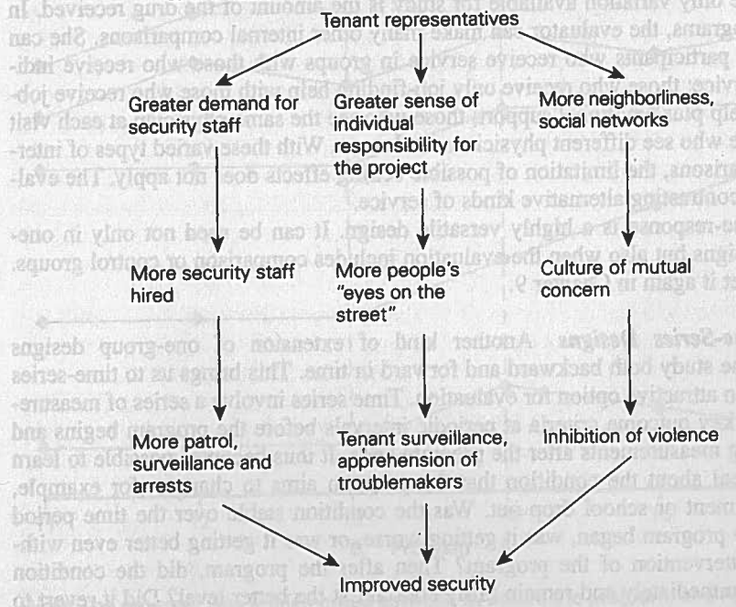
ory makes about the course of the program's development. To the extent that things work out as posited in the program's theory of change, the evaluation has grounds for confidence that it understands not only what is going on but how program effects are taking shape as well.

When the evaluator designs the study around program theory, she chooses the key nodes of the theory as the content and time points for data collection. For example, a program of tenant participation in the management of a public housing project assumes that such participation will improve resident security. The link between tenant representation and improved security isn't obvious, but the program theory hypothesizes that it will come about through one or more of the claims illustrated in Figure 8-3.

The evaluator then designs methods for collecting data about the extent to which each of these assumptions actually comes to pass. She locates appropriate sources and sets appropriate time intervals. If one or more of the assumptions are borne out, she has a better sense of how the program is working.

Dose-Response Designs With one-group designs, there is no comparison of participant outcomes with people who did not receive the program. Sometimes such comparison is impossible because every eligible person is served. This is the case with Social Security, public schooling, hospital emergency rooms, meat and poultry

FIGURE 8-3 PROGRAM THEORY MODEL: THE ASSUMED LINKS BETWEEN TENANT PARTICIPATION IN THE MANAGEMENT OF A PUBLIC HOUSING PROJECT AND RESIDENT SECURITY



inspection, Medicare, and many other programs. In such cases, an *internal* comparison can usually be designed. The evaluation can compare participants who received a large quantity of service (a high dose) with those who received a lower dose. The assumption is that, if the program is doing good, more of it will do better. The evaluation examines whether high-dose recipients have better outcomes than their low-dose counterparts.

The notion that more is better may not always be right. One can imagine a program that gives service so sufficient that more of it would not prod outcomes up a single notch. It would reach a ceiling, and a comparison with higher levels of services would unfairly suggest that it was not effective. But such a situation would be rare. Most programs are constrained in resources and struggle to offer service to all those who are eligible. They do not provide *all* the help that they know how to and want to give. On the other hand, if there is reason to believe that a program has approached its ceiling, it might be possible for the evaluator to arrange a lower service level cohort for comparison. The question that such a study would be answering is: Would a reduced level of service, at lower cost, reach equally good results?

With a dose-response design, the evaluator has to take care that those who receive more service are not different on other grounds, too, that they are not more conscientious in attendance, or have higher motivation, or are judged more in need of service by staff. More conscientiousness or motivation would be likely to lead to better outcomes in and of themselves; greater need for service would probably predict poor outcomes. But if the amount of service received is uncorrelated with such other characteristics, the internal comparison can be revealing.

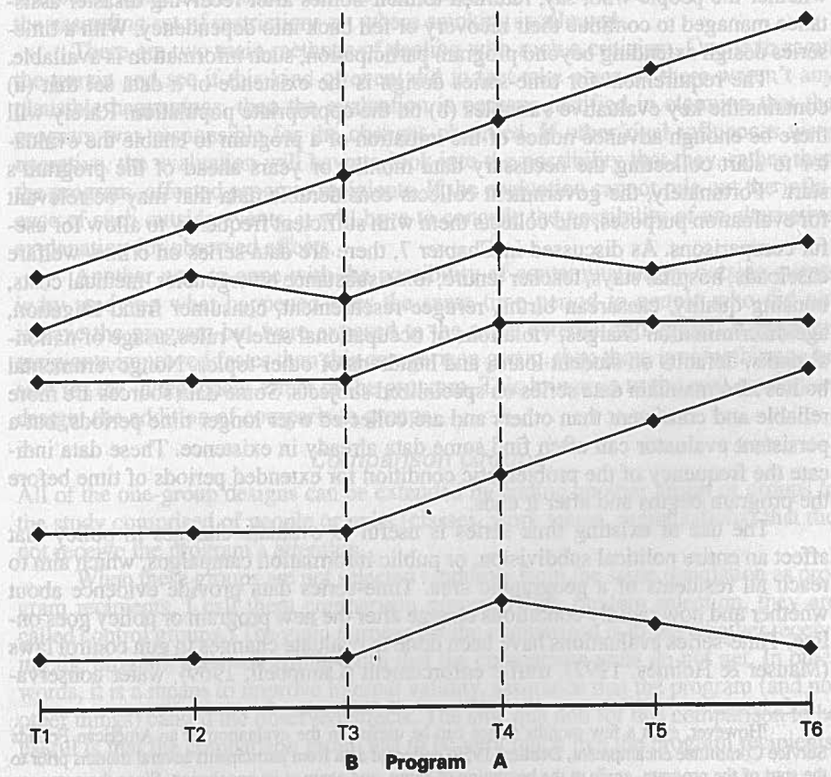
The term *dose response* comes from research on drugs, and in that circumstance the only variation available for study is the amount of the drug received. In social programs, the evaluator can make many other internal comparisons. She can compare: participants who receive service in groups with those who receive individual service; those who receive only job-finding help with those who receive job-finding help plus emotional support; those who see the same physician at each visit with those who see different physicians; and so on. With these varied types of internal comparisons, the limitation of possible ceiling effects does not apply. The evaluation is contrasting alternative kinds of service.

Dose-response is a highly versatile design. It can be used not only in one-group designs but also when the evaluation includes comparison or control groups. We'll meet it again in Chapter 9.

Time-Series Designs Another kind of extension of one-group designs extends the study both backward and forward in time. This brings us to time-series designs, an attractive option for evaluation. Time series involve a series of measurements on key outcome criteria at periodic intervals before the program begins and continuing measurements after the program ends. It thus becomes possible to learn a great deal about the condition that the program aims to change—for example, unemployment or school drop out. Was the condition stable over the time period before the program began, was it getting worse, or was it getting better even without the intervention of the program? Then after the program, did the condition improve immediately and remain fairly constant at the better level? Did it revert to its original problem status as effects of the program faded out? Or did good results escalate over time, with success generating further success?

Time-series data enable the evaluator to interpret the pre-to-post changes in the light of additional evidence. They show whether the measures immediately before and after the program are a continuation of earlier patterns or whether they mark a decisive change. Figure 8-4 shows five cases in which the same degree of change occurred over the span of the program: The slope of the line from before (B) to after (A) the program is the same. But in these cases the change has different significance. The top line shows a case where things were already improving, and the program doesn't appear to have speeded or slowed down the rate of change. In the second case, the condition is erratic. Improvement seems to occur between B and A, but it is just the same kind of jiggle that happened before without a program and

FIGURE 8-4 FIVE CASES OF TIME-SERIES DATA MEASURING PROGRAM EFFECTS AT SIX TIME POINTS



Note: The program was implemented between B, before the program, and A, after.

happened again some time after the program had ended. It is not obvious that the program made any difference.

In the third case, the change from B to A appears to be the result of the program. Things were going along at a consistent level, then escalated after the programs and remained at the higher level. In the fourth case, outcomes improved after the program and continued to improve thereafter. Perhaps the program set off a train of events (improved self-insight, community mobilization) that led to continued development. The final case shows improved outcomes following the program, but the results fade out after the program ends. If the program wants to sustain participants at the higher level of functioning, it will have to institute further intervention.

Time-series data address the question of persistence of effects, an important but often neglected question. Rarely do evaluations go on long enough to follow participants for more than a year or so after their departure from the program. Even when the evaluator plans such long-term follow-up, evaluation funds can run out or be cut off before the long-term data are collected. So there is little information about whether the people who, say, returned to their homes after receiving disaster assistance managed to continue their recovery or fell back into dependency. With a time-series design extending beyond program participation, such information is available.

The requirement for time-series design is the existence of a data set that (a) contains the key evaluative variables (b) on the appropriate population. Rarely will there be enough advance notice of the initiation of a program to enable the evaluator to start collecting the necessary data months or years ahead of the program's start.[7] Fortunately, the government collects considerable data that may be relevant for evaluation purposes, and collects them with sufficient frequency to allow for useful comparisons. As discussed in Chapter 7, there are data series on crime, welfare caseloads, hospital stays, teacher tenure, toxic substance emergencies, medical costs, housing quality, caesarean births, refugee resettlement, consumer fraud litigation, age discrimination charges, violations of occupational safety rules, usage of national parks, defaults on student loans, and hundreds of other topics. Nongovernmental bodies also maintain data series on specialized subjects. Some data sources are more reliable and consistent than others and are collected over longer time periods, but a persistent evaluator can often find some data already in existence. These data indicate the frequency of the problematic condition for extended periods of time before the program begins and after it ends.

The use of existing time series is useful to evaluate changes in policy that affect an entire political subdivision, or public information campaigns, which aim to reach all residents of a geographic area. Time-series data provide evidence about whether and how rapidly conditions change after the new program or policy goes online. Time-series evaluations have been done to evaluate changes in gun control laws (Mauser & Holmes, 1992), traffic enforcement (Campbell, 1969), water conservation strategies (Maki, Hoffman, & Berk, 1978), seat belt use (Desai & You, 1992; Rock, 1992), and sale of lottery tickets (Reynolds & West, 1978).

[7]However, even a few months notice can be useful. In the evaluation of an American Friends Service Committee encampment, Dentler (1959) collected data from participants several months prior to the start of the program, again at the beginning of camp, and again at its conclusion. Since the camp ran for less than two months, the additional data point helped to answer questions about whether the camp had an influence or whether the youth were changing anyway.

Where no data are available that fit the study's needs, the evaluator can *begin* the collection of relevant data. Obviously such a step will not fill in earlier history. "Pre-pre-program" data can be collected only with early warning and long lead time. But the evaluator can plan for "after-after-after" data—that is, for continuing the periodic collection of outcome data for a long period. She makes her respondents into a panel for continued reinterviewing.[8] Panel data fulfill the same functions as time-series data. The main differences are that panel data, being tailor made for the evaluation, are apt to be a more exact match to the needs of the evaluation but, being started for the sake of the evaluation, are apt to be shorter in duration.

The main criticism of the validity of evaluations that rely on time-series (and panel) data is that some outside events coincided with the program and were responsible for whatever change was observed between before and after, or between times 1, 2, 3, and 4. For example, it wasn't necessarily the program that accounted for the decline in smoking; it could have been the television series on the risks of smoking that came along at the same time or the publicity given to a new medical study or the cascading set of restrictions on where smoking is allowed.

There are two main methods of dealing with such a criticism. One is to scout the terrain and see if this kind of event did in fact take place. If there weren't any plausible happenings, then the evaluation is perhaps justified in claiming that the program was responsible for the changes observed. If other rival influences *were* operative, the evaluation will have to look into the possibility that they, rather than the program, affected program recipients. If the evaluation cannot rule out the influence of such outside events, it will have to concede the possibility of an alternative explanation for observed effects.

Another way to cope with the possibility of contamination by outside events is by studying what happened over the same time period to people who did not receive the program but were exposed to the same external influences. If program recipients improved faster than this comparison group, then there is something to be said for the independent effect of the program. This brings us to the next section on design: the addition of comparison groups.

Comparison Groups

All of the one-group designs can be extended by adding another group or groups to the study comprised of people or units (classes, work teams, organizations) that did not receive the program's attention.

When these groups are not selected randomly from the same population as program recipients, I call them comparison groups. (With random selection, they are called control groups.) The main purpose for the comparison is to see whether receiving the program adds something extra that the comparison units do not get. In other words, it is a means to improve internal validity, assurance that the program (and not other things) caused the observed effects. The sine qua non for this comparison to be useful is that the comparison group has to be very much like the program recipients.

[8]Where the evaluator collects data from her own panel, the data are usually not called time-series data but panel data.

At the end of the study period, they are going to be the surrogates for what the program recipients *would have been like* if they had not been in the program.

After-Only with Comparison Group The after-only design can be strengthened by adding a comparison group that is as similar to program recipients as possible. Matching people in the program to similar people who have not participated in the program is a common method of constructing a comparison group in after-only designs. Studies have used next-door neighbors of participants, their older or younger siblings, residents of the same neighborhoods, students in the same schools. In the after-only design, the evaluator has only posttest measures for participants, and she is not sure what their status was at the start of the program. Nor is she sure that next-door neighbors or students in other classrooms were like them at the outset on characteristics that matter. The attributes on which she is likely to match are standard demographic variables, such as age, race, and sex, because these characteristics do not change over the course of the program and they are the easiest to get data about. Sometimes these factors matter. But the evaluator often has scant reason for expecting that they are the ones most likely to affect outcomes.

When records are available on relevant items for the before period (welfare status, school achievement, days of absence from work), these will be eminently useful to fill in the missing before information. If key items can be retrieved for both the program group and the comparison group, this design begins to approximate the "before-and-after with comparison group" design. However, it is unusual for an evaluation to locate all the relevant information in existing records. When factors that influence success in the program are intangibles like motivation or density of social support networks, the likelihood of finding *before* information drops close to nil.

Still, the addition of a comparison group helps to strengthen the ability to make causal inferences. Suppose the recipients of the program do well on an outcome measure (say, released prison inmates are arrested in much lower numbers than expected in the year following the program). Was this showing the result of the program or something else? Adding a group of released inmates who are similar in many ways helps to answer the question. If an equivalent proportion of them are staying clear of the law, perhaps the program shouldn't get the credit. But if they are recidivating at a much higher rate than program recipients, the suggestion that the program is responsible gains credibility.

The comparison group will almost inevitably differ from the participant group in important ways. The sheer fact that participants selected themselves, or were selected by others, into the program is a persuasive indication. Often the evaluator doesn't know which variables are likely to affect outcomes and so doesn't know what kind of comparison group to recruit. Without pretest data, although perhaps with makeshift pretest data, it is especially difficult to disentangle the prior differences from the effects of program service. The best course is usually to extend data collection and move to before-and-after design.

Before-and-After with Comparison Group This is one of the most commonly used designs in evaluation. The comparison group is selected to be as much like the client group as possible through any of a variety of procedures. But it is not

randomly assigned from the same population, as would be a true control group in an experimental design.

Often the evaluator can locate a similar site (preschool, university department) serving people much like those in the program that did not operate the program. Then she can compare the status of both groups before the program began and test how similar they were. If there are strong similarities in the pretest on such items as age, socioeconomic status, or whatever items are relevant to the program's purpose, *and* on items assumed to be predictive of success on the outcome indicators, such as motivation and skill, she can proceed with some confidence.

But of course nothing is quite so easy. The comparison group is likely to be different from the program group. For one thing, the program group often selected themselves into the program. They were motivated enough to volunteer. Or staff selected them into the program because of their special need for its services or because of judgments that they would particularly profit from it. People who choose to enter a program are likely to be different from those who do not, and the prior differences (in interest, aspiration, values, initiative, even in the neighborhood in which they live) make postprogram comparisons between served and unserved groups problematic. In job training programs, evaluators have constructed comparison groups from lists of unemployed persons who would have been eligible for the program but had not had contact with it. Others have used unemployed friends of job trainees for comparison. Other evaluators have used people who registered for the program but did not enter (no shows) or those who dropped out (Bell, Orr, Blomquist, & Cain, 1995). Checking often discloses that the groups are not comparable in several respects.

The search for controls who are as like program participants as possible has led to the use of unawares (people who did not hear of the program but might have joined had they heard) and geographic ineligibles (people with characteristics similar to participants but who lived in locations that had no program). Comparison groups have been used in such studies as the evaluation of the supplemental nutrition program for pregnant women, infants, and children (WIC) (Devaney, Bilheimer, & Schore, 1991) and the National School Lunch and Breakfast Programs (Burghardt, Gordon, Chapman, Gleason, & Fraker, 1993).

Each ingenious strategy solves some problems and raises others. What was there about the unawares that blocked their knowledge of the program? What are the effects of community conditions in the different location? The question has been raised whether it is important to eliminate self-selection bias in program evaluation. Since voluntary programs inevitably include self-selected participants, would it be appropriate to evaluate the combined effects of self-selection and program participation? Such a procedure would certainly simplify the comparison group problem and the evaluator's life. Study results would apply to the combination of volunteering and receiving services. It would not be obvious how much of the effect was attributable to the program per se. But if future programs are also going to be available on a voluntary basis, the evaluation results will provide useful information.

Sometimes it is useful to use several comparison groups (students in another community college who are like the program group in the type of educational institution attended and students in a state university who are like the program group in

location). Each of these comparisons will shed light on one feature, and each will compensate for differences that the other comparison leaves uncontrolled.

At the least, the evaluator should consider whether program recipients started out as better or poorer risks than the comparison group. Often it is the better risks who were selected or selected themselves in. If they show better outcomes than the comparison group, a plausible explanation is that they would have prospered whether there was a program or not. On the other hand, if the program attracted those most in need, the comparison group started out with an advantage. Even if the program does an admirable job, the program group may not catch up. The change will have to be larger than the initial gap between the two groups before the program registers any effect at all. A more privileged comparison group tends to minimize the observed effects of the program. Initial differences have to be taken into account in analysis and reporting of the data.

Multiple Time Series A comparison can also be added to the time-series design. If the evaluator can find a similar group or institution and locate periodic measurements of it over the same time span as the program group, she can learn a great deal. This design appears particularly appropriate to evaluations of school programs, since repeated testing goes on normally, and a long series of pre- and postscores are often available.

A well-known example of multiple time series was the evaluation of the Connecticut crackdown on highway speeding. Evaluators collected reports of traffic fatalities for several periods before and after the new program went into effect. They found that fatalities went down after police began strict enforcement of penalties for speeding, but since the series had had an unstable up-and-down pattern for many years, it was not certain that the drop was due to the program. They then compared the statistics with time-series data from four neighboring states, where there had been no changes in traffic enforcement. Those states registered no equivalent drop in fatalities. The comparison lent credence to the conclusion that the crackdown had had some effect (Campbell, 1969; Campbell & Ross, 1968).

Multiple time series are especially useful for evaluating policy changes. When a jurisdiction passes a law altering fire codes, policing practices, or taxation on business, there are opportunities to compare the trends in fire damage, arrest rates, and bankruptcies. Without necessarily mounting a whole new study, officials can examine conditions over a period of time before the new policy was adopted and for a period of time after it went into force, and then make comparisons with similar jurisdictions that did not experience the policy change. One of the advantages of this design is that the localities do not have to be very similar at the outset on many dimensions so long as they were similar on their trends in fire damage, arrest rates, or bankruptcies.

Time-series designs also lend themselves to a number of multiple comparisons. If data are routinely collected, it is probable that they are routinely collected in many places, often under the aegis of higher level governments—new claims for unemployment insurance, admissions to drug rehabilitation programs, and so on. Therefore, the evaluation can draw upon multiple comparisons. Some localities will be similar to the community running the program in size, others will be similar in

socioeconomic composition, others in degree of urbanization, and so on. The study can examine how the program community compares to other communities that are similar on different dimensions. It may also become possible to analyze which dimensions are related to the trends in outcomes that are observed. With several series of data, comparisons can take account of these dimensions, singly and in combination, in interpreting program outcomes.

Constructing a Comparison Group A nonequivalent comparison group should be as similar to the program group as human ingenuity can manage. (If the human were ingenious enough to manage random assignment of program and nonprogram groups from the same population, she would have a control group.) Without random selection and assignment, she often resorts to matching. She can try to find a match for each person who is entering the program group on all the variables that are likely to affect that person's performance on outcome measures. For example, in a program that provides self-care instruction for hospital patients, she can try to find a "comparison someone" who is similar to the Asian 45-year-old diabetic woman who receives the program. But matching one to one is often difficult to do. In our heterogeneous polyglot society, there is not a ready match for every program participant. If the evaluator cannot locate a match for the Asian 45-year-old diabetic, she may have to drop the woman from the study. This leads to a loss of cases and a lower sample size, a condition that is bad for the power of statistical comparisons. Losing cases also makes the evaluation less representative of the whole program population. The study will not be generalizable to those like the participants for whom matches could not be found.

Without trying for a one-to-one match, the evaluator can try to find a group of patients whose overall profile is similar to the program group. Afterward, when one group has been exposed to the benefits of the program and the other group has not, the difference in the gain each group has made over the same time period is an estimate of the effect of the program.

Sometimes this is the best that can be done. The comparison group is added to one-group designs or time-series designs to rule out the likelihood that maturation or outside events or simply becoming adept at filling out evaluation protocols was responsible for whatever changes were observed. But matching is less satisfactory than randomized assignment on several counts. Not the least is that the evaluator often cannot define the characteristics on which people should be matched. She doesn't know which characteristics will affect whether the person benefits from the program. She may have matched on age, sex, race, and IQ, when the important factor is parental encouragement. Some wit has said that if we knew the key characteristics for matching, we wouldn't need the study.

In job training programs, where a great deal of evaluation has been done, evaluators have learned that prior work experience is a critical variable. In programs that aim to move families out of poverty, family composition (e.g., the presence of two adults or working-age children) is an important predictor. In education programs, test scores depend to a considerable extent on whether the teacher has covered in class the material that is included on the test. Even when knowledge is available about which variables are implicated in success, data are not always available that

allow evaluators to match on that basis. No existing data show which small entrepreneurs have willing family members to call on, and so, however important, this feature cannot become the basis for matching.

Another problem is finding people with the right constellation of characteristics. The woods are not necessarily full of Latino college applicants who are receiving scholarship support. Locating matches for some program groups is complicated by their uniqueness, by the availability of similar services to other people in their situation, and by the perennial difficulty of enlisting the support of comparison group members to whom the program gives nothing. (This latter problem afflicts random-assignment controls as well. The evaluation asks them to cooperate in providing data periodically over a sometimes lengthy period of time, without getting anything in return. Small wonder that attrition rates are sometimes high.)

Matching is sometimes done on the basis of pretest scores. The program serves students who are consistently doing poorly in school. To find a comparison group, the evaluator looks at test scores in other classrooms and selects students whose test scores are very low. This tends to be a poor procedure.[9] It can produce particularly misleading results when program participants and comparisons are drawn from basically different populations. Regression to the mean gets in the way of valid interpretation. Regression to the mean is not intuitively obvious, but it is based on the fact that a person's pretest and posttest measures are not perfectly correlated even without a planned intervention of any kind. Statistically, posttest measures are expected to be closer to the group mean. At a second testing, what look like effects of the program may be artifacts of statistical regression. For example, Figure 8-5 shows a case in which there are no program effects. Each group has simply regressed to its mean. An unwary evaluator may ascribe the shift to the effects of the program when in fact there are no program effects at all.

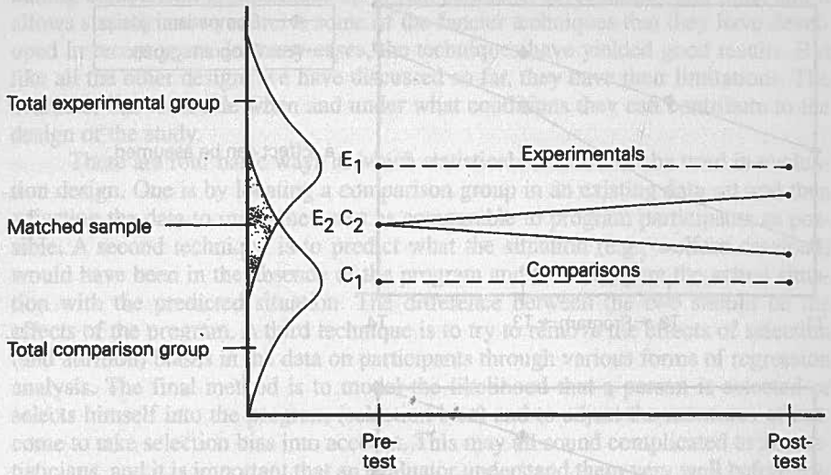
Regression to the mean can be compounded by measurement error. All measures contain some component of error, and some, such as test scores and attitude measures, contain a sizable amount. On any one testing, some individuals will score artificially high and others artificially low; on a second testing, their scores are likely to be closer to their usual score.

It is better in dealing with nonequivalent controls to compare the measures of natural groups than to select only extreme cases by matching. In Figure 8-5, this would mean using the E_1 and C_1 measures rather than E_2 and C_2.

Comparison groups created by matching can reduce two of the main threats to the validity of study conclusions—outside events and maturation—but they are not an adequate control for selection. As for attrition, when members of either the program or comparison group drop out, the matched pair can be eliminated from the analysis, but in so doing two unhappy events occur. As with the failure to find initial matches in the first place, sample size and subsequent statistical power are reduced, sometimes drastically, and the sample becomes progressively less representative of the total program population. It is usually a better strategy to exploit all

[9]Note that we are talking about matching without randomization. If units are matched and then randomly assigned to each group, the procedure increases the statistical precision of the experiment. Matching as a prelude to randomization may even be essential when there are few units, such as cities. But matching as a substitute for randomization can produce pseudoeffects.

FIGURE 8-5 REGRESSION TO THE MEAN: REGRESSION ARTIFACTS IN A MATCHED SAMPLE



the available data gathered prior to dropout and adjust for differences between program recipients and comparisons that appear at the end.

When members of the comparison group are not very similar to the program group, some inferences can be made if pretest data are available. The evaluator looks at the growth of participants and nonparticipants (Moffitt, 1991). If participants improve at a steeper rate than comparisons, even if their initial status was different, the data may suggest a program effect. But such an inference is uneasy.

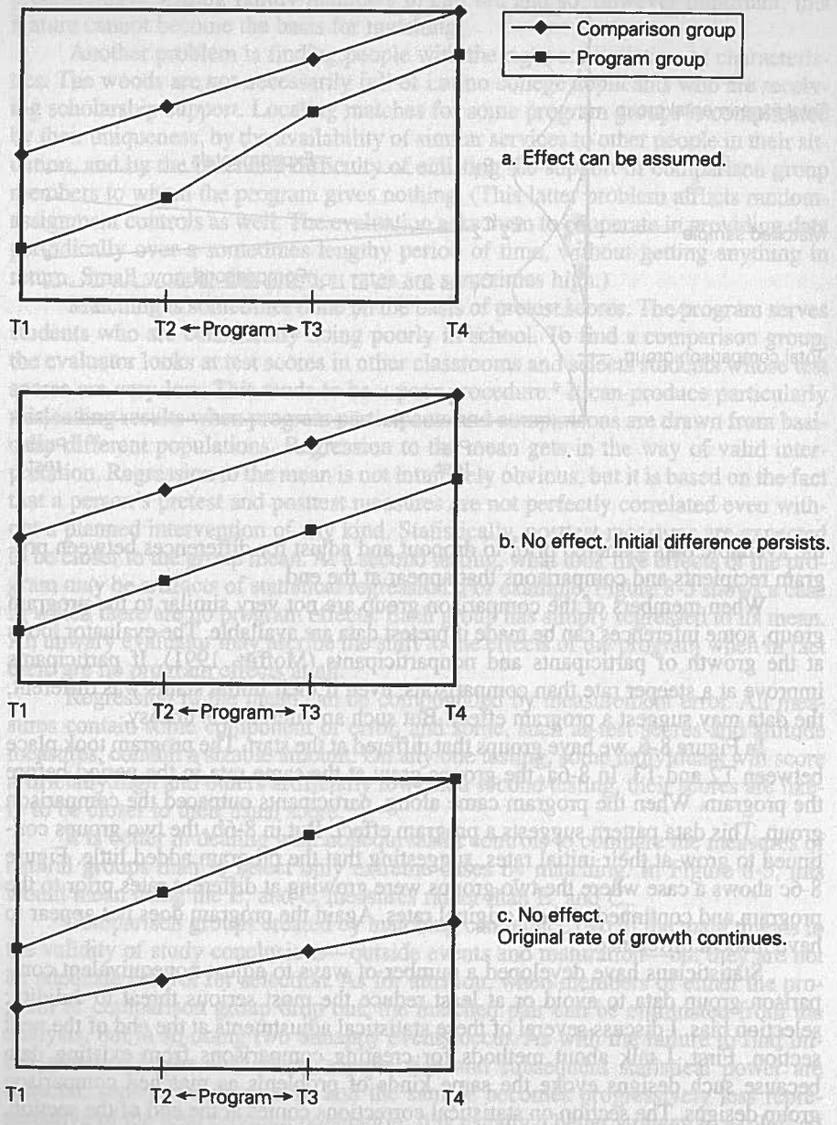
In Figure 8-6, we have groups that differed at the start. The program took place between T2 and T3. In 8-6a, the groups grew at the same rate in the period before the program. When the program came along, participants outpaced the comparison group. This data pattern suggests a program effect. But in 8-6b, the two groups continued to grow at their initial rates, suggesting that the program added little. Figure 8-6c shows a case where the two groups were growing at different rates prior to the program and continued at their original rates. Again the program does not appear to have made a difference.

Statisticians have developed a number of ways to adjust nonequivalent comparison group data to avoid or at least reduce the most serious threat to validity: selection bias. I discuss several of these statistical adjustments at the end of the next section. First, I talk about methods for creating comparisons from existing data because such designs evoke the same kinds of problems as matched comparison group designs. The section on statistical corrections comes at the end of the section.

Statistical Techniques in Lieu of Matching

So far we have talked about constructing a comparison group by matching program participants to nonparticipants who share many of the same characteristics. The problem has been where to find similar people and how to ensure that they are as

FIGURE 8-6 THREE TIME-SERIES DESIGNS WITH COMPARISON GROUPS



similar as possible to the participants at the beginning of the program. In large-scale evaluations, matching is often supplanted by use of available data, which are then adjusted statistically. The procedure follows the same logic as matching but makes use of data already collected and adjusts the data to match the participant group.

The use of existing data avoids the necessity of finding, recruiting, and maintaining comparison groups made up of real people. It saves money and time, and it allows statisticians to exercise some of the fancier techniques that they have developed in recent years. In many cases, the techniques have yielded good results. But like all the other designs we have discussed so far, they have their limitations. The evaluator has to decide when and under what conditions they can contribute to the design of the study.

There are four basic ways in which statistical methods can be used in evaluation design. One is by locating a comparison group in an existing data set and then adjusting the data to make the cases as comparable to program participants as possible. A second technique is to predict what the situation (e.g., welfare caseload) would have been in the absence of the program and then compare the actual situation with the predicted situation. The difference between the two should be the effects of the program. A third technique is to try to remove the effects of selection (and attrition) biases in the data on participants through various forms of regression analysis. The final method is to model the likelihood that a person is selected or selects himself into the program (selection bias) and to adjust the measures of outcome to take selection bias into account. This may all sound complicated to nonstatisticians, and it is important that an evaluator understand them very well before trying to put them to use. But the logic is fairly easy to grasp.

Comparison Group in an Existing Data Set In its simplest form, this involves finding a data source that has information on a population roughly similar to the participants in the program. The data may come from such longitudinal series as the National Longitudinal Survey on Youth, Panel Study on Income Dynamics, or High School and Beyond. Evaluations of job training programs often rely on the Census Bureau's Current Population Survey, which has large national samples of individuals, or data from Social Security records (Ashenfelter & Card, 1985; Bloom, 1987; Bryant & Rupp, 1987; Dickinson, Johnson, & West, 1987). Chapter 7 lists a number of national longitudinal data series that have useful data for comparisons.

Once having located a data set, the evaluator should check (a) the population that was surveyed, to see how similar it is to program participants, and (b) the questions asked, to see whether they ask for the information that the evaluation requires in a form that fits the study. If the evaluator has consulted existing surveys early in the game, she may find questions and measures that are suitable for her evaluation, and she can use the identical items to ensure comparability.

When a survey is satisfactory on counts of both population and measures, the evaluator can use data from the total sample as a comparison. Even though the population is not very close to program participants in all respects, the data provide an overall comparison that throws light on the condition of people who did not receive program service.

The evaluator may next want to extract a subset of respondents from the data set who are more similar to the recipients of the program than is the total sample. That means that she has to get the raw data file and manipulate it. Problems can arise even after locating and obtaining the data set: The data are not transferable to the evaluator's computer system without major work in converting and reformatting.

The original database may not be well documented, so that it is not clear what the items mean or how they were collected and coded. The data may be aggregated at a level that doesn't match the unit of analysis being used in the evaluation. Available computer software may be unable to operate on the data as they are formatted (Stewart & Kamins, 1993).

Moreover, data that come from existing data sets are not fresh from the field. They had to be collected months or even years earlier because it takes time to collect, check, code, process, clean, organize, and document the data, and then prepare data tapes or disks for public use. By the time the evaluator receives the data set, it may be several years out of date. Whether or not this is a matter of concern depends on the stability of the information over time and the nature of the comparison the evaluator wants to make.

Many of the large data sets now come on CD-ROMs, which provide quick, accurate, and relatively inexpensive access to the information. They require a CD-ROM player and software designed to retrieve the data on the disk. However, new technology is in the offing, and systems may change relatively soon. They are probably going to keep changing, as electronic wizardry continues to develop.

If the evaluator still wants to go ahead and select a subsample from a relevant data set, she has to identify the variable(s) on which she will choose the subsample, such as education or employment status. She will seek a variable or a very few variables that are associated with the program outcome.

Let us take a dropout prevention program for high school students. The evaluator knows from previous studies that variables associated with dropping out of school include socioeconomic status of students' families (low), school grades (low), and age in grade (older than the average student in that class). She looks at a large national data set and finds individuals with a wide range of values on these variables. She can attempt to select individuals in the data set who are close matches on one or more of these variables. Since she is dealing with several waves of data on the same individuals, she has to decide which year's data to use for selecting the subsample. Unless the evaluator is skilled in analysis of large longitudinal surveys, she should get expert help.

Even when everything goes well, the cases in the dataset subsample may still differ from the program group in their overall distribution. For example, 40% of the dataset cases may show a grade point average lower than 2.0, compared to 55% among program participants. Statistical weights can be introduced to give more weight to each case in the underrepresented category, or differences between participants and comparisons can be reduced through the use of statistical adjustment. See the section on statistical adjustments below.

With all the technical and conceptual problems, the use of existing data to form an artificial comparison group has much to recommend it. Obviously, it saves the time, expense, and psychic wear and tear of recruiting real people to receive no program. It also appeals to the sense of frugality. Here are all these multi-million-dollar longitudinal surveys with opulent quantities of excellent data, collected with conscientiousness and creativity by highly talented researchers available almost for the asking. It feels wasteful to ignore them. Moreover, the datasets provide information on large numbers of people and help to answer a recurrent evaluation question: How do changes in program participants compare to changes in other people in roughly similar circumstances over the same time interval?

Probably the easiest way to make good use of the information is to look at the distribution of responses for the *total survey population* (often from published materials) on data items central to the evaluation over analogous periods of time. Where the survey population differs from the program population, the evaluator should point out the differences and estimate how the differences affect the comparison. For example, if the survey respondents on average come from higher income families, on the basis of past research they can be expected to show faster gains on reading test scores than the program population. The comparison, therefore, will exaggerate the gap between the two groups and tend to underestimate the growth in reading scores of people in the program. When published data, or easily accessible tabulations, show survey data separately by family income, then program participants can be compared to the subgroup in the survey whose incomes are most similar to theirs.

Forecasting the Situation without the Program Another way that statistical techniques can be used is to extrapolate past trends into the future, as a way of estimating the counterfactual—that is, the condition that would have occurred if no program had been operating. Thus, for example, an evaluation of a dropout prevention program can use past school data to project the rate of future school dropouts, on the assumption that past trends continue into the future. Then the evaluator compares the actual rate of school dropouts with the projected rate at given points in time and assumes that any observed difference is due to the program.

A more sophisticated form of this procedure was used to evaluate a nutrition assistance program in Puerto Rico (Beebout & Grossman, 1985). Data were taken from two food intake surveys carried out in 1977 and 1984. Changes in the food stamp caseload were modeled before the program was implemented. After the program, the evaluators compared current estimates of food expenditures to the expenditures that would have occurred, based on the model, if there had been no program. Similarly, Garasky (1990) projected the welfare caseload in Massachusetts based on data from 1976 to 1983, when a state employment and training (ET) program went into effect.

In another study (Holder & Blose, 1988), simulation was used to model the effects of a complex of factors on alcohol-related traffic accidents in a community. The simulation took into account such factors as the consumption of alcohol in the population by age, sex, and drinking practices, vehicle miles driven, legal age limit for purchase of alcohol, enforcement of driving-under-the-influence statutes and conviction rates, and disposable income. The effects of each of these factors on alcohol-related traffic accidents was modeled on the basis of the best available research. The simulation yielded projections of the frequency of injuries and fatalities for the years ahead under a range of different assumptions—all assuming the absence of any programmatic intervention.[10] Once an intervention was implemented, projected

[10]Such a simulation model could be used to estimate the effects of an intervention before it is implemented in order to figure out whether it is likely to make a positive difference. Computer simulations can be a planning tool (Holder & Blose, 1988).

injuries and fatalities could be compared to actual injuries and fatalities, with the simulation in effect providing the comparison group.

One of the key limitations to this approach is one that we have met before: outside events. Events outside the program that are not accounted for in the statistical model can be responsible for changes in welfare cases or food expenditures or dropouts. For dropouts, for example, availability or unavailability of jobs might lead to changes in leaving school. Or changes within the schools may be involved, such as alterations in curriculum, school administration, or teaching staff. This procedure can be supplemented with other techniques, such as comparison groups, to try to identify and deal with changes due to outside events.

Statistical Adjustments for Preexisting Differences between Program Participants and the Comparison Group Whether a comparison group is constructed of similar people or from data in existing surveys, their nonequivalence to the program population is a constant affliction. This is selection bias, which we have met before.

A common method for seeking to equate program participants and members of nonrandom comparison groups is to control statistically for the variables on which they differ. The idea is simple: Identify the variables that are likely to affect program outcomes and then remove the differences between the groups on those variables by analysis. For example, if participants are more likely than the constructed comparison group to be new arrivals in the company at the start of the executive training program, control for length of time in company.

Regression analysis is one of the statistical techniques that can be used for this purpose. In effect, it controls for differences between the program group and the comparison group. The variables may be measures of health status, income, length of employment, severity of crime committed, or whatever else there is reason to believe differentiates the groups. Regression analysis can be used to estimate the extent to which each of these variables predicts program outcomes, and once these have been taken into account, differences in the outcome that remain are assumed to be due to participation in the program. It is a method for equalizing the two groups on those factors that the evaluator knows about, has measured, and has entered into the analysis.

In multiple regression, large numbers of factors can be controlled simultaneously, and additional terms can control for the interactions among them. If the evaluator has identified and appropriately measured the relevant differences between the two groups, the end result is to leave the two groups much the same for analytic purposes. Any differences in outcomes can then be ascribed to the effects of the program.

That is the general idea of statistical adjustment. In practice, more complex procedures are usually used. The problem of noncomparable groups does not yield readily to solutions, and statisticians incorporate a variety of procedures into their quest for comparability. The key is good knowledge or theory about which variables matter for gauging the effects of the program.

Many studies have used these kinds of techniques (e.g., Ashenfelter, 1978; Dobson, Grayson, Marshall, O'Toole, Leeder, & Schureck, 1996; Kiefer, 1979; Lee & Loeb, 1995), either alone or in conjunction with other statistical procedures, such as selecting comparisons from existing datasets. Controlling for differences through statistical controls has proved to be highly useful. However, a recurrent limitation is the inability to account for *all* the factors that distinguish the two groups. Those who do and do not enter a program differ in a variety of ways, not all of which are known or measured. Still, with increasing experience and with increasingly sophisticated statistical methods, much progress has been made. In some fields such as job training, hundreds of evaluations have been conducted, and a good deal of knowledge has accumulated. Increasingly, complex statistical methods have been introduced, and continued discussions and debates among analysts lead to further refinements in method.

Modeling the Selection Process One of the currently popular analytic refinements has been to model the selection process itself. That is, the evaluator develops a regression equation that includes all the elements presumed to affect the decision to enter the program. These become the independent variables. The dependent variable is participation in the program. The solution to the equation is a predicted probability that each case will be in the program group. This prediction score for each individual is then entered as a control variable into the overall regression (Heckman, 1980; Heckman & Hotz, 1989).

Many evaluators have adopted the technique. It has staunch supporters but critics as well. Its value rests on good knowledge of the characteristics that are predictive of participation in the program and on the availability of good data that measure those characteristics. Inventive statisticians have developed other correction techniques. Donald Rubin, for example, has developed a procedure using what he calls propensity scores (Rosenbaum & Rubin, 1983, 1984).

Before we leave this section, let me note that the evaluator who attempts to use these statistical methods has to have a full understanding of the techniques and of the assumptions on which they are based. She also needs the technical training to use appropriate statistics wisely. In addition, she has to understand the measures that were used in the evaluation, their statistical properties, and distribution. Consultants can solve some of the problems, but it is important for the evaluator to understand what the consultant is doing. Fortunately, a good statistician can often explain the logic of the procedures in terms that are accessible to people with a modest level of statistical expertise.

One final note: An issue of the *Journal of Educational Statistics* (Wainer, 1989) dealt with problems of interpretation created by nonrandom samples, and the authors suggested several solutions. After pages of text, equations, graphs, and unexpected sprightly prose, the organizer of the issue, statistician Howard Wainer, listed a set of conclusions. Among them were the following:

> Statistics can't work magic.
>
> Without information, the best we can do is to characterize our ignorance honestly.
>
> It is important to know the subject area so that we can better model selection.
>
> The only consensus we can hope for among competing analysts is the kind of testing any proposed adjustment strategy has to successfully undergo prior to acceptance. (p. 188)

The testing continues.

When to Use Nonequivalent Control Group Designs

Now that I've discussed methods for constructing and adjusting comparison groups, let's consider when it is beneficial to rely on them. Obviously, there are times when the evaluator has little choice. It's either a nonequivalent (nonrandomized) comparison group or no comparison at all. Evaluators agree that some comparison is better than none. The more similar the comparison group is in its recruitment and the more similar their characteristics on pretest measures, the more effective they will be as controls. To repeat, any differences between the groups should be measured and reported, and the evaluator should indicate the likely direction of bias stemming from noncomparability, noting whether it tends to underestimate or overestimate program effects.

Comparisons with nonequivalent groups are useful, but several studies urge that such comparisons be used with caution. Evaluators have studied cases where both nonrandom comparison groups and randomized control groups have been available. When they compared the results, they found that nonrandom comparisons often gave misleading results. This has been true even after statistical adjustments were introduced to control for known differences between the groups. Boruch and his colleagues reviewed results where experimental evidence on program effects was collected alongside evidence from quasi-experiments or econometric modeling. They conclude on the basis of a series of studies that the procedures can yield results that are very wide of the mark (Dennis & Boruch, 1989; National Research Council, 1989). Using data from evaluations of training programs, two sets of investigators (Fraker & Maynard, 1987; Friedlander & Robins, 1994; La Londe, 1986; La Londe & Maynard, 1987) compared the results of randomized experimentation, quasi-experimentation, and econometric modeling. They relied on the randomized experiment for providing the true estimate of program effects; in contrast, the other procedures ranged from overestimating effects by 50% to underestimating them by 2,000%. Similarly, evaluations of the Salk polio vaccine were undertaken by both randomized experiments and by an early quasi-experiment in which second graders received the Salk vaccine and first and third graders were used as the comparison group. The quasi-experimental results differed considerably from those of the experiment (Dennis & Boruch, 1989). Shadish and his colleagues compared randomized and nonrandomized evaluations of marital and family therapy (Shadish & Ragsdale, 1996) and education and drug abuse prevention programs (Heinsman & Shadish, 1996). They found sizable differences in results, but when crucial design features were taken into account, nonrandomized studies more closely approximated randomized results. Further comparisons of randomized and nonrandomized designs will be useful in understanding the conditions under which nonrandomized designs produce results that are similar to those of experiments.

In conclusion, let me recommend an important strategy. If the evaluator can determine the selection of even a fraction of the program recipients, she can assign them randomly to program and control conditions. At least for this subset of recipients, she will be able to make valid estimates of the difference that program participation makes.

Summary

This chapter reviews a variety of research designs for evaluating the process and the outcomes of social programs. Informative data can be obtained by asking people on the scene about their experiences. Experts can be called in to render judgments based on their experience and knowledge of programs of the same type.

These sources offer useful information, particularly on inputs and processes of program implementation. When the evaluation is meant to apprise directors, sponsors, and funders about outcomes, the evaluator will want to use designs that are less vulnerable to bias and incompleteness.

In a more formal mode, the evaluator can collect outcome measures on the program group *after* the program has been in operation and seek to impute participants' status prior to entry. This design is likely to be used when the evaluator is called in too late to collect *before* data herself. Better is a before-and-after design, where comparable data are collected pre- and postprogram. The design can be strengthened by adding *during* measures while the program is in midstream. Following the tracks of program theory improves the design further by investigating whether the hypothesized links between process and outcome materialize.

Time-series designs add additional data points. Starting long before the program begins, continuing "during-during-during" the program, and going on for several years after the program ends, time series indicate whether the pattern of outcome measures supports the conclusion that the program is responsible for observed effects. Time-series data also show whether positive effects are sustained over time, accelerate, or fade away. Multiple time series add comparisons with sites that did not receive the program. Such comparisons are a way to rule out the threat that outside events, rather than the program, caused observed outcomes.

Existing longitudinal data series are an excellent source of time-series data. Thousands of such series are available at local, state, and national levels. Where no source covers the topics required for the evaluation, the evaluator can seek to collect appropriate data over time by repeated interviewing of panels of respondents. The longer the interviewing can continue, the more suitable will the data be for time-series analysis.

One-group before-and-after designs can be strengthened by the addition of comparison groups who do not receive the program. One way to create comparison groups is by matching nonparticipants to participants on characteristics that are related to desired outcomes. Thus, the evaluator often seeks people of the same age, race/ethnicity, gender, socioeconomic status, severity of condition, or other key attributes. In order to avoid differences in other unmeasured attributes, such as motivation, the evaluator will try to find people who had not heard about the program, who live in communities where the program was unavailable, or who were on the waiting list.

Evaluators can create artificial comparison groups by using available data sets containing similar people who were not served by the program. They can try to select out the subset of respondents most similar to program participants for comparative purposes. Before-and-after comparisons with these respondents offer clues about whether exposure to the program was the key factor in change.

Statistical methods can be used to try to equate participants and comparisons, whether the comparison groups are developed by matching procedures or from existing datasets. The most common statistical techniques involve controlling for variables that are expected to affect outcomes. When their effects on outcomes are accounted for, any remaining difference between participants and comparisons is assumed to be the result of participation in the program. A cautionary note: The effect of any variables that are not measured and not included in the analysis is not controlled. Since evaluators often cannot identify all the variables that influence outcomes, and since they do not have measures of some variables even when they can identify them, statistical corrections can lead to under- or overestimating program effects. Another statistical procedure is to identify the factors that differentiate those who enter the program from those who do not and to model the selection process. When this can be validly done, the effects of selection bias can be controlled.

The purpose of these increasingly complex designs is to rule out the possibility that things other than the program are causing whatever outcomes are observed. They seek to counter threats to the validity of the causal conclusion and improve internal validity. In evaluation a number of threats to validity have been identified, such as maturation, outside events, testing, and instrumentation, but the most pervasive and serious problem is selection: Those who enter a program are usually unlike those who do not enter on a number of measured and unmeasured characteristics.

Evidence about which variables are implicated in desired outcomes comes from research and prior evaluations, and better knowledge is accruing over time. Measurement of the appropriate variables depends on the state of the measurement art and the availability of well-measured variables in existing datasets. Measurement progresses, but with cutbacks in federally supported data series, less data are likely to be available and intervals between collection points are lengthening. Statistical techniques are advancing to cope with discrepancies between program and comparison groups, and more studies are comparing the results of nonequivalent comparison groups and random-assignment control groups. From these efforts, evaluators are getting a better understanding of which techniques work well to equate experimental and comparison groups and where further progress is required.

Although the informal and quasi-experimental designs in this chapter do not control against all threats to causal attribution, many of them provide sufficient information to size up the situation. Where serious ambiguities remain about the causal link between the program and observed outcomes, the evaluator can patch on further investigation. The key is to recognize what the weaknesses are and to compensate for them through parallel—or additional—inquiry.