CHAPTER 6

Evaluating Outcomes

CHAPTER OUTLINE

- ***The evidence-based paradigm.*** Recently, we have seen widespread adoption of a new normative value: that criminal and juvenile justice systems should implement programs that have been proven through rigorous evaluation research to be effective.
- ***Types of evaluation.*** Impact, performance, and efficiency. Is the program or policy achieving its objectives, are outcomes changing over time, and is it worth the investment of resources devoted to its implementation? Impact evaluations depend on methods that enable us to attribute observed outcomes to the intervention, rather than other potential influences. Additionally, meta-analyses combine the results of many individual evaluation studies to determine if different intervention approaches or practices work better than others.
- ***Three prerequisites for evaluation must be met.*** (1) Objectives must be clearly defined and measurable, (2) the intervention must be sufficiently well designed, and (3) the intervention must be well implemented.
- ***Evaluability assessment and logic modeling.*** These are two methods for examining the components of a program in preparation for evaluation. Both methods also help clarify program and policy designs.
- ***Outcome measures should be derived from an intervention's objectives.*** Good outcome measures are *valid* and *reliable* indicators of results in terms of specific program aims.
- ***Identify potential confounding factors*** (factors other than the intervention that may have biased observed outcomes). Common confounding factors include biased selection, biased attrition, and history. There are three major techniques for minimizing confounding effects: (1) random assignment, (2) nonequivalent comparison groups, and (3) propensity score analysis.

- ***Specify the research design to be used.*** Examples include: the simple pre-post design; the pre-post design with a control group; the pre-post design with multiple pretests; the longitudinal design with treatment and control groups; the cohort design; and time series analysis.
- ***Identify users and uses of evaluation results.*** Who is the intended audience, and how can results be effectively and efficiently communicated? How will the results be used?

Now the time has come to measure the impact of the intervention: Has the program or policy achieved its intended effect(s)? How can we tell? The goal at this stage is to develop a research design for measuring program or policy outcomes (a specific, intended change in the problem, defined by objectives). Did the program or policy achieve its intended objectives? Why or why not?

In spite of how obvious the need for evaluation may seem, many programs and policies have never been evaluated. Cost is often given as a reason for not conducting evaluations, but we must also recognize that evaluation can be threatening to stakeholders because their public image, political power, and/or agency budget is linked to the success or failure of a specific program or policy. Sometimes not evaluating is an effective means of maintaining the status quo without accountability for performance.

Increasingly, however, funding agencies are demanding accountability for outcomes. Grants made to public and private agencies by federal agencies such as the National Institute of Justice typically require an evaluation component, and often an independent researcher must do the evaluation. In the fields of health and mental health, managed care agencies carefully measure outcomes and costs in order to ensure that the money they manage is being used effectively. This kind of thinking is making its way into the criminal justice system.

Remember, when we construct objectives, we identify a result and a criterion (standard for measurement) for each objective (Chapter 3). These objectives become the focus of evaluation.

THE EVIDENCE-BASED PARADIGM

Recently, we have seen widespread adoption of a new normative value: that criminal and juvenile justice systems should adopt programs that have been proven through rigorous evaluation research to be effective. This process of replicating proven program designs implies that existing programs that are ineffective are being replaced or radically re-engineered. The term "evidence-based" is now part of the language

of policymaking, and several organizations are working hard to promote the use of these proven program designs, particularly in the areas of juvenile justice and behavioral health. These sources include:

- *CrimeSolutions.gov* is an online, searchable source of program effectiveness information in such areas as crime prevention, substance abuse, and juveniles. This service of the National Institute of Justice contains up-to-date information as well as detailed information on programs that have been evaluated. In keeping with the work of Mark Lipsey (Lipsey & Wilson, 2001), www.crimesolutions.gov also documents practices that effective programs have in common. Programs are classified as effective, promising, or ineffective.
- *National Center for Mental Health and Juvenile Justice* (NCMHJJ): Evidence-based practices are defined as interventions that involve standardized treatment and that have been shown through controlled evaluation research to produce improved outcomes across multiple research groups. Evidence-based practices suggested by NCMHJJ include: Multisystemic Therapy (MST), Brief Strategic Family Therapy (BSFT), and Therapeutic Foster care (see http://www.ncmhjj.com).
- *Blueprints for Violence Prevention* (BVP): BVP, a project developed by the University of Colorado at Boulder, with funding from the Colorado Division of Criminal Justice, Centers for Disease Control and Prevention, the Pennsylvania Commission on Crime and Delinquency, and Office of Juvenile Justice and Delinquency Prevention (OJJDP), has researched programs that are considered "model programs" as well as "promising programs" based on the following criteria: evidence of deterrent effect with a strong research design, sustained effects, and multiple-site replication. Programs identified by BVP include Multisystemic Therapy (MST), Functional Family Therapy (FFT), and Aggression Replacement Therapy (ART) (see http://www.colorado.edu/cspv/blueprints).
- *OJJDP: The Model Programs Guide* (MPG): This guide was developed by OJJDP to assist practitioners and communities in implementing evidence-based prevention and implementation programs. OJJDP uses a rating system based on the evaluation of the literature of specific prevention and intervention programs based on four criteria: the conceptual framework of the program; program fidelity; the evaluation design; and empirical evidence demonstrating the prevention or reduction of problem behavior, the reduction of risk factors related to problem behavior, or the enhancement of protective factors related to problem behavior. Model programs include: Cognitive Behavioral Treatment (CBT), Functional Family Therapy (FFT), and Wraparound Case Management (see http://dsgonline.com/mpg2.5/mpg_index.htm).

TYPES OF EVALUATION

Before we define what kind of evaluation data to collect, it is important to know about different approaches to evaluation. First, we need to be clear about the kinds of evaluative questions we are going to ask before designing the evaluation. If we want to know whether a policy is achieving its objectives, an impact evaluation is needed. However, evaluations of programs and policies can take one of three major approaches: (1) impact assessment, (2) continuous evaluation, or (3) efficiency analysis. Another type of evaluation, process evaluation, focuses on whether a program was implemented as it was designed. Process evaluations are not intended to measure program effectiveness.

Note that we do not intend (or pretend) in this chapter to cover evaluation methods in all their complexity (see Berk & Rossi, 1998; Patton, 1996; Rossi et al., 2003; Wholey et al., 1994). We do intend that readers should become familiar with some basic concepts necessary to understand evaluation. We will not attempt to teach students or practitioners how to design their own measures in this book; that is a task for a good course in research methods.

Impact Evaluation

The most common type of evaluation, and the type we focus on mainly in this chapter, is an *impact evaluation.* To assess impact, we want to compare *actual* outcomes to *desired* outcomes (objectives). In order to do this, we will need valid measures of the desired outcomes and information about the status of clients on these measures prior to their exposure to the intervention. For example, the fact that a high proportion of clients of a delinquency prevention program end the program with high self-esteem is meaningless if they started the program with high self-esteem. Moreover, it is not sufficient to know simply that a change occurred: we need to determine whether the program or policy in question caused the observed change. We need to know that this change wouldn't have happened without the intervention. To know this, we will also need information on similar types of persons who were not exposed to the intervention. If the same change occurred in this second group, then we are unable to attribute the change to the intervention.

We may also be interested in long-term effects of a program. Recidivism, for example, is not something we measure prior to the program, but is likely to be influenced by changes that the program is designed to bring about, such as increased attachments to others, improved anger control, or academic achievement. In order to assess the impact of a program on a long-term outcome, we need a comparison group of persons that did not receive the program. If recidivism among those who attended the program is no different from the comparison group, and the two groups are identical in terms of those characteristics believed to be associated with recidivism, then we have evidence that the program does not work.

Performance Evaluation: Outcome-Based Information Systems

One weakness of many impact evaluations is that their results are limited to a specific point in time. Programs, in particular, are constantly changing in terms of their staff, clients, services, and goals. Staff turnover, intervention fads, changes in the political environment, and changes in the characteristics of incoming clients can all produce changes in program outcomes. The results of even the best-designed impact evaluation gradually become obsolete. *Performance evaluation* offers an alternative: why not collect and analyze outcome information on all clients on a permanent basis? This way, stakeholders could learn from the outcome data, make adjustments, and see the consequences of their responses over time. The growth of computerized information systems within criminal justice is making this approach increasingly viable and useful.

This incremental learning process incorporates much of what we talked about in the last chapter with regard to monitoring, but the focus is on improving outcomes rather than assessing the adequacy of implementation. Moreover, this interactive approach to evaluation incorporates the concept of "action research" that was introduced by Kurt Lewin (see Chapter 2). In terms of design, an outcome-based information system is a *multiple cohort design* (see the section "Specify the Research Design" in this chapter) in which the outcomes for each cohort (a specific group of clients) can be compared as a trend over time.

Let's add one more step to this idea: such a system of learning is even stronger when outcome information is monitored for an entire system of programs (Oberweis et al., 2004). For example, a city or state could monitor specific outcome data on all programs that serve a specific population of clients, such as drug offenders or juvenile delinquents. The comparison group is not made up of clients who receive no services but very similar clients (a matched comparison group) who receive different services.

With this information, comparisons of programs over time generate more information about why certain outcomes are being produced. If an intervention is showing positive gains, consideration should be given to conducting an evaluation using an experimental design. Computerized information systems are increasingly common within criminal justice. Although the focus of such systems is typically on management needs—for example, personnel, finance, and case control—client-specific outcome information can easily be added to enhance the capacity of the organization to assess a program or policy's success.

Efficiency Evaluations

Lastly, we may want to know how efficient a given program or policy is. Two types of analyses are useful for this purpose: cost-benefit and cost-effectiveness analyses. In *cost-benefit analysis,* we ask if the amount of change that is being produced (the

benefit) is worth the cost (usually in monetary terms). *Cost-effectiveness analyses,* in contrast, express outcomes in substantive terms so as to compare programs and policies that produce similar outcomes. In this case we are viewing programs or policies as competitors, using both cost and outcomes as criteria for judging relative benefit.

For example, let's consider a three-month school-based curriculum for teaching problem-solving skills to eighth graders determined to be at high risk for delinquency. If desired changes are taking place, then various measures of benefit can be established (e.g., a unit increase in problem-solving skills as measured by a standardized test results in a decrease in the number of children who avoid any subsequent arrest up to age 18). The costs, some of which may be less obvious than others, must now be accurately measured. For example, how much money did it cost to buy course materials (notebooks, videos, etc.) and train classroom teachers to administer the curriculum? Was there a cost in terms of what students *didn't* get (e.g., a reduction in mathematics or science training to free up space for the problem-solving skills curriculum)? We may then compare the difference between total dollars saved by preventing each arrest (e.g., costs of arresting, processing, charging, and supervising each adjudicated delinquent), and dollars expended on operating the problem-solving skills curriculum. In other words, we calculate ratios of costs to benefits. A moderate gain at a low cost may often signal a more worthwhile program than a slightly higher gain at a much more substantial cost. A cost-effectiveness analysis of the same program might estimate the total dollars spent to convert a cohort of delinquents into nondelinquents, and then compare this approach with others that attempt to produce similar outcomes (e.g., secure detention). We can then judge which program is most efficient in producing desired outcomes.

We can also compare actual costs against projected costs: an "efficient" program might be one that came in on budget or under budget, but produced a tangible benefit. In many cases, the costs of achieving the same level of benefit can be compared across different programs, policies, or even settings. Efficiency analyses, therefore, can provide valuable information to assist stakeholders and policymakers in making choices from among competing programs, policies, and projects.

Exactly how specific costs and benefits should be defined, however, is a matter of some controversy, and procedures for conducting efficiency analyses tend to be quite complex. The Vera Institute of Justice has developed a toolkit available for free on the basics of cost-benefit analysis (Henrichson & Rinaldi, 2014), largely because we recognize the power of cost-related arguments with policymakers (http://www.vera.org/pubs/cost-benefit-analysis-justice-policy-toolkit). For all three types of evaluation, we expect our readers to be aware of why, how, and where such analyses are used, but their actual conduct requires sophisticated training and expertise. This is particularly true of efficiency analyses (Stokey & Zeckhauser, 1978; Thompson, 1980).

Meta-Analysis

Although an impact evaluation can tell us about the success or failure of a particular program, we can't be sure what components of the program are causing the observed outcomes. Most programs provide several discrete activities, such as group counseling, recreation, behavioral contracting, and education. How can we be sure which of these activities or which combination of activities is causing success? Of course, programs are more than activities: there are other clients attending the program, the staff, the facility, and the management of the organization providing the program. All of these program characteristics can affect a client's experience in a program, and consequently affect program outcomes.

In order to discover whether certain treatment approaches work, researchers like Mark Lipsey (Lipsey & Wilson, 1993, 2001) and Paul Gendreau (Gendreau & Little, 1996) have made use of a statistical method that allows us to examine alternative causes of program outcomes. By selecting only those evaluations that are well designed, meaning that the study used an experimental or suitable quasi-experimental design, meta-analysis involves calculating the size of the difference between those who received the intervention and those who did not in each study, called an "effect size," and then combining these effect sizes across different studies into one measure of the impact of the treatment approach (Lipsey & Wilson, 2001).

To date, meta-analyses have found that behavioral and cognitive-behavioral methods are more effective for delinquent youths than various types of client-centered, nondirective therapies (Lipsey & Wilson, 2001). For adult correctional treatment, cognitive-behavioral and behavioral approaches work better than other treatments. Intensive prison drug treatment appears to be effective, especially when combined with community aftercare (Lipsey & Cullen, 2007). Education, vocational training, and prison labor programs have modest effects. The effects of sex offender treatment are uncertain. Findings of meta-analyses, we caution, are often limited by methodological weaknesses of the studies analyzed (e.g., selection bias), a lack of detailed information about the subjects and the treatment, and/or questionable implementation or program fidelity (Gaes et al., 1999; Welsh & Zajac, 2004).

THREE PREREQUISITES FOR EVALUATION

Before actually evaluating a program or policy, three main criteria (prerequisites) must be satisfied. These prerequisites are defined in Figure 6.1. If any one of these prerequisites is not met, any attempt at evaluation is likely to be unsuccessful, and the results will be unconvincing. Indeed, an entire methodology called "evaluability assessment" (Rossi et al., 2003; Rutman, 1984; Wholey et al., 1994) has been developed to address these critical concerns.

1. Program or policy objectives must have been clearly specified, and those objectives must be measurable (see Chapter 2).

2. The intervention must have been well designed (see Chapter 3). Its logic should be meaningful and easy to comprehend.

3. The intervention must have been well implemented so that there is no question that its critical elements (activities) have been delivered to clients as planned. Remember, this is why you do monitoring: to find out whether the program or policy in action matches the program or policy on paper (see Chapter 5).

FIGURE 6.1 *Three Prerequisites for Evaluation*

EVALUABILITY ASSESSMENT

Under those circumstances in which we are intending to evaluate an existing intervention, it is often advisable to first examine the program or policy to determine what aspects of the intervention can be measured. It may be that a program has clearly articulated goals but has not completed the task of creating measurable objectives. If one program goal is to see an improvement in offenders' perceptions of their opportunities for the future, how will the program know if these perceptions have changed? It may also be the case that different program staff members have different opinions regarding the program's goals. To the extent that these differences exist, the evaluation may be targeting the wrong objectives.

An evaluability assessment is a method for uncovering actual program components and isolating those elements that can be measured (Wholey, 1994). This evaluable model of the program is the program that is tested in the evaluation. Any program elements that cannot be measured are put aside. The assessment typically involves reading written documents that describe the program, interviewing administrators and line staff members, reading case files, and even interviewing program clients. Each of these sources of information is queried as to program goals and objectives, program activities, impact models, program resources, planning mechanisms, and areas in which change is needed. The result of the evaluability study is a single program model that is then reviewed by program administrators and staff for accuracy. The model often forces program personnel to confront their differences and for the first time achieve consensus regarding important facets of their program.

LOGIC MODELING

You will recall from Chapter 2 that an *impact model* links an intervention to the causes of a problem, and the causes of the problem to the problem itself (see Figure 6.2).

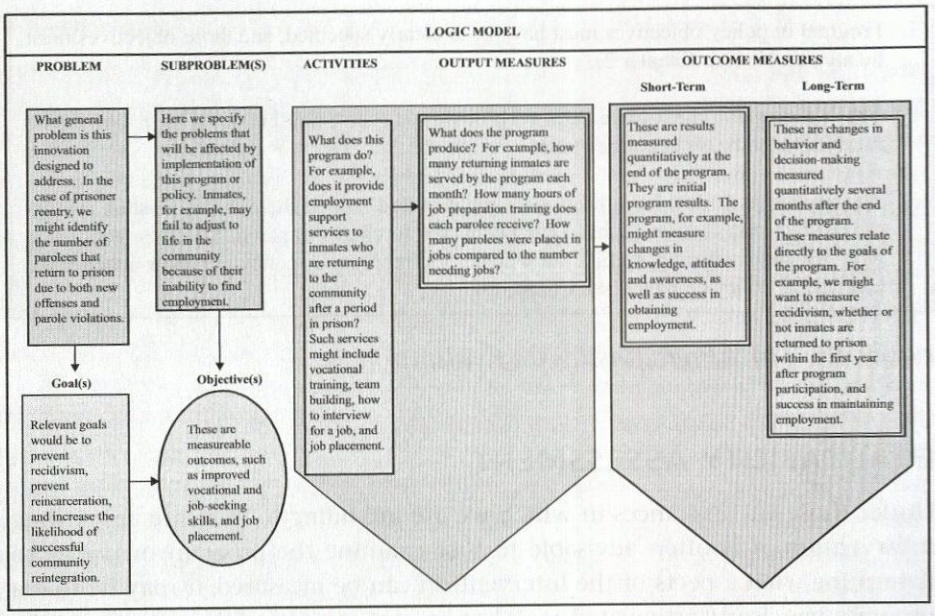

FIGURE 6.2 *Logic Model*

The logic of the model is that the intervention will change the causes of the problem in order to affect the problem. When we evaluate interventions, it is often useful to map out the logic of the program or policy. These maps, which are actually drawn on a page, are called *logic models.* Logic models are a concise way to see how a program is designed.

The logic model summarizes the main components of the program and shows how the activities of the program are related both to the program's objectives and to measurable outcomes (Forgatch et al., 2005; Guevara & Solomon, 2009; Mihalic, 2004). An example of a logic model is shown in Figure 6.2. A logic model can be used to communicate effectively with stakeholders the intent of the program or policy and its underlying strategy, and it can provide a framework for designing an evaluation. It makes clear the outcomes that should and can be measured in order to test the effectiveness of the intervention.

Why Develop a Logic Model?

A logic model can be used for at least four different purposes:

1. *Strategic and Program Planning*—Developing a logic model is a form of strategic planning. The process forces you to identify your vision, the rationale behind your program, and how your program will work. This process

is also a good way to get a variety of program stakeholders involved in program planning and to build consensus on the program's design and operations.

2. *Effective Communications*–Logic models allow you to provide a snapshot view of your program and intended outcomes to funders, staff, policymakers, the media, or other colleagues. They are particularly useful for funding proposals as a way to show that what you are doing is strategic, and that you have a plan for being accountable.

3. *Evaluation Planning*–A logic model provides the basic framework for an evaluation. It identifies the short-term and ultimate outcomes you are aiming for–based on your program's design–and puts those outcomes in measurable terms.

4. *Continuous Learning and Improvement*–A completed logic model provides a point of reference against which progress toward achievement of desired outcomes can be measured on an ongoing basis.

IMPLEMENTATION ASSESSMENT

A second assessment is often conducted prior to the evaluation study itself in order to be sure that the program or policy being evaluated has been implemented with fidelity to the design, that is, that their activities and practices conform with the design, that dosage levels are appropriate, that the staff are competent to deliver services, or that decision makers are trained regarding procedures and provisions of a policy (the template for this logic model was retrieved July 20, 2015, from the OJJDP web site at: http://ojjdp.ncjrs.gov/grantees/pm/logic_models.html). We are mainly interested in the impact of an intervention that has been designed well. If during or following implementation aspects of the program or policy are changed, it will be impossible to make sense of outcomes that are measured and analyzed. It is critical, then, that evaluators ensure that the program or policy has been implemented by program staff or by decision makers as it was designed.

DEVELOPING OUTCOME MEASURES

To develop outcome measures, refer back to the intervention's objectives. How will we adequately measure these objectives? Remember that an adequate objective contains four components (see Chapter 2): time frame, target population, a key result, and a criterion for measurement. We are trying to determine whether a specific intervention (program or policy) produces an intended change in the problem. Recall from Chapter 3 that an impact model specifies such a prediction or hypothesis.

Establishing the impact of a program amounts to establishing causality. In other words, we want to determine whether the intervention produces a specific effect, an intended change in the problem. To do so, we need adequate measures, and an adequate research design.

The *validity* of a measure refers to the degree to which any measure or procedure succeeds in doing (measuring) what it purports to do. Most experts refer to this type of validity as *construct validity*. In other words, how can you tell whether your measure actually assesses the construct or concept that it is supposed to assess?

For example, we might be using a measure of self-esteem, such as the Rosenberg Self-Esteem Scale (Rosenberg, 1965), or we might be using a measure of self-reported drug use, such as the National Survey on Drug Use and Health (available at http://www.oas.samhsa.gov/nhsda.htm). The question is: How accurate is each measure? Is it a good indicator of the construct you are trying to measure? These questions are often investigated through research that attempts to demonstrate that the measure relates to some known indicator of the same concept. We might measure a student's self-esteem, and then correlate self-ratings with ratings from that student's friends and family members. We want to see if there is a relationship between our measure and some other indicator of the same concept. Or, we might validate self-reported drug use with actual drug-testing technology to determine if self-reports are under- or over-inflated. Wherever possible, we try to use existing measures for which previous research has indicated reasonable evidence of validity.

The *reliability* of a measure refers to its consistency. For example, what is the probability of obtaining the same results upon repeated use of the same measuring instrument (i.e., test-retest reliability)? We want to be sure that the measure is somewhat consistent over time, and that results don't vary dramatically from one time to the next. For example, self-esteem is seen as a relatively stable personality trait. Any reliable measure should not yield wildly disparate results about a person's self-esteem from one week to the next. Attempts to establish reliability of a self-report measure such as self-esteem usually examine, through research, the internal consistency of the items in a measure (i.e., do items correlate with one another), or relationships between scores obtained from two or more separate administrations of the same test.

IDENTIFYING POTENTIAL CONFOUNDING FACTORS

Establishing the impact of a program, as we noted above, amounts to an attempt to establish causality. Did the intervention produce an observed change in the problem? Before we look at a few basic research designs, we need to discuss *confounding factors* (sometimes called confounds). These refer to any factors, *other than your program,* that may account for observed changes on the outcome measure (e.g., an increase or decrease in the problem). Confounding factors bias the measurement

of program outcomes. In research design textbooks, these confounding factors are often labeled threats to the internal validity of the experiment.

Biased selection is one common confound of which to be careful. In many criminal justice interventions, especially offender treatment and post-release programs, researchers view reducing recidivism as a primary objective. However, upon close inspection of many interventions, we often find out that many of the clients who were selected to receive the treatment weren't high risk to begin with. If youths in a delinquency prevention program had no observable risk factors at the start—such as previous arrests, truancy, academic failure, or family problems—it is not surprising if such youths, upon graduation from the program, show a low rate of recidivism. Does this mean that the intervention worked? Or does it mean that client selection was so biased that we have no way of knowing whether the program actually works?

When we refer to confounding factors, we are saying that something else (other than the intervention itself) may have caused the observed change in the problem, or something may have disrupted (confounded) the way we measured a change in the problem. Confounding factors introduce bias into our measurement of outcomes. You need to anticipate potential confounds and design your evaluation to minimize them. The evaluation of each and every intervention should address potential confounds. Here are three of the most typical confounds:

1. *Biased selection:* Systematic bias in client selection procedures results in the treatment group not including adequate numbers of clients with demonstrated needs or problems. Sometimes called "creaming," this problem occurs when a program deliberately or unknowingly selects those clients most likely to show a favorable outcome, rather than those clients most in need of the intervention. For example, many private drug treatment programs claim phenomenal rates of success, but we often find that they have limited their client selection to those with the least severe problems. In other words, clients most in need were not selected, and our suspicions are further aroused if there was no control group against which to compare program outcomes. We have no faith whatsoever in such results.

2. *Biased attrition:* Bias is introduced into the outcome measure because subjects dropped out of one comparison group at higher rates than subjects in other comparison groups. For example, it is a common difficulty in drug treatment programs that those with the most severe problems drop out before the end of the program. The observed result is that those who remained in the treatment program had lower rates of relapse than similar subjects in a control group. The result is biased, however, because the treatment program lost those subjects who were most likely to show the highest rates of relapse.

3. *History:* Some unanticipated event, occurring between the beginning and the end of the intervention, introduces bias into the measurement of program objectives. For example, if a major change in a state's law regarding domestic violence occurred during the course of a mandatory arrest experiment, the new law, rather than the intervention, might explain the observed result of increased arrests for spousal abuse.

Example 6.1 Potential Confounds in the Minneapolis Domestic Violence Experiment

The Minneapolis Domestic Violence Experiment (Sherman & Berk, 1984) has been criticized for potential confounding factors. Some suggest that we cannot adequately determine from this experiment whether a mandatory arrest policy works better than mediation or separation. One measure used was a follow-up interview with victims to ask about victimization following the police intervention. Victims were interviewed immediately after the intervention, then every two weeks for 24 weeks. Researchers reported a decrease in the problem, as measured by fewer victim reports of repeat abuse. Of couples who received the mandatory arrest intervention, only 19 percent of victims reported further abuse in the follow-up study, compared to repeat abuse rates of 33 percent and 37 percent, respectively, in the separation and mediation interventions.

Here is the difficulty: what if women were scared to report further incidents of abuse because they had been threatened or beaten by their spouse following the previous police intervention (arrest, mediation, or separation)? Sherman and Berk reported that a substantial number of victims in their sample dropped out of the study. Initial interviews with victims were completed in only 62 percent of all cases. Others couldn't be found, or they refused to be interviewed. Biweekly interviews were completed for only 49 percent of subjects in the original sample. The study may have lost many of those who were victims of repeat abuse following the experiment.

Sherman and Berk reported that of those victims who they actually contacted, there was no "differential" attrition (i.e., the victim dropout rate for the experiment was about the same for each of the three interventions) (Sherman & Berk, 1984). However, we have no way of knowing how many of those not contacted actually experienced further abuse. Critics expressed doubts about the experimental results because of this potential confound (Lempert, 1989). In addition, attempts to replicate the results of the Minneapolis experiment in other jurisdictions have not been very successful (Sherman, 1992).

In summary, it is not entirely clear that the intervention (mandatory arrest) was responsible for the observed results (a decrease in reported incidents of abuse). We cannot completely rule out the possibility that the results were biased due to the attrition of more than half of the original subjects. We are suspicious that those victims who refused to be interviewed in the follow-up study might have been more likely to experience further victimization than those who agreed to be interviewed. The observed reduction in repeated incidents of abuse may be due to the fact that victims who dropped out of the experiment were afraid to report further incidents of abuse to police or interviewers.

MAJOR TECHNIQUES FOR MINIMIZING CONFOUNDING EFFECTS

There are three major techniques for minimizing confounding effects: (1) random assignment, (2) nonequivalent comparison groups, and (3) propensity score analysis. Each involves creating a comparison group: a different group of clients that is equivalent to the treatment group on any factors that might influence the outcome measure, such as recidivism, but doesn't receive the intervention.

Random Assignment

Random assignment means that researchers randomly assign eligible clients to separate treatment and control groups. This is not the same as random selection, which would make absolutely no sense whatsoever. People often have a hard time keeping these two concepts separate. As a sampling strategy, one might randomly select subjects to participate in a survey or opinion poll. The purpose would be to obtain a representative, random sample of the *general* population. In contrast, nobody would ever randomly select clients for an intervention; they would instead determine who is eligible for the program, and who needs the program. Once the *eligible* pool of clients is determined, they might then randomly assign subjects to the treatment and control groups. Clients in any intervention are not randomly selected; they are deliberately selected on the basis of need and eligibility. Once selected, they might be randomly assigned to treatment or control groups.

What random assignment does, in theory, is equalize two different groups on unknown differences (e.g., intelligence, previous criminal history, etc.) that might bias the outcome results. With a large enough sample, the chances of equally distributing characteristics of subjects across the treatment and control groups is very good. This is the best method for dealing with confounds, when possible. For ethical and practical reasons, however, random assignment is not always possible. In many social interventions, those most in need of the intervention must be selected, and randomization would be unfair or even unethical (see the example at the end of this chapter).

Nonequivalent Comparison Groups

Often we cannot randomly assign subjects to treatment and control groups, but we can attempt to construct treatment and comparison groups made up of clients with similar characteristics. It is especially important that the two groups are similar in terms of their level of need, and in terms of characteristics that might influence the outcome of interest (e.g., recidivism). We must decide with care exactly which factors might be important to control for. We usually look to previous research to determine important variables. We might then attempt to create matched control groups, so that average client characteristics are distributed relatively equally across the two groups (aggregate matching), or so that every client in the treatment group

is matched one-to-one with a similar nonclient in the control group (individual matching). Aggregate matching is much easier than individual matching, unless one is dealing with an extremely large pool of eligible clients. Individual matching is more precise, but we lose a very large number of potential cases as we try to match individuals rather than groups on a large number of variables. For example, we may be measuring recidivism as a program outcome. It would be important to match our treatment and control groups on variables known to influence recidivism, such as previous criminal behavior, age of offender, substance abuse, job skills, and so on.

Propensity Score Analysis

Recent evaluation studies have illustrated that it is possible to identify comparison groups when random assignment is not possible, using methods that control for naturally occurring differences between the treatment and comparison groups. A method known as *propensity scoring* allows researchers to estimate treatment effects by collapsing variables known to be associated with the decision to place individuals into the program into a single score, and controlling for that score (Rosenbaum & Rubin, 1983, 1985; Rubin, 1997). This creates comparison and treatment groups that can be considered comparable on all known variables that are believed to be related to the outcomes of interest. Propensity score analysis approximates a randomized control trial by matching subjects on a single score that represents all known factors that may have been associated with their being assigned to the program being evaluated.

SPECIFY THE RESEARCH DESIGN

In this section, we attempt to acquaint readers with a few of the most basic research designs used to evaluate program impacts. In general, such designs specify when and how measures will be collected to assess program impact. Each involves comparisons of certain groups of subjects, and measurement on specific variables, over particular time periods, to evaluate outcome. We will diagram and describe several of the most commonly used research designs.

You might want to think of an example as you go through the diagrams and descriptions of different research designs. Imagine, for example, a six-week prevention program designed for adolescents at high risk of abusing drugs. The program attempts to raise youths' self-esteem. The program's rationale is that increasing self-esteem is a means of increasing one's ability to make independent decisions without being unduly influenced by one's peers. In order to measure change in self-esteem, a self-reporting self-esteem instrument such as the Rosenberg Self-Esteem Scale is used (Braga et al., 2008; Gover et al., 2003). This change is a short-term outcome. The long-term outcome would be drug use at a later point in time.

In each of the diagrams below:

"X" represents the intervention or treatment

"O" represents the observation or measure

"PRE" refers to a pre-intervention observation

"POST" refers to a post-treatment observation

FIGURE 6.3 *Legend for Diagrams of Research Designs*

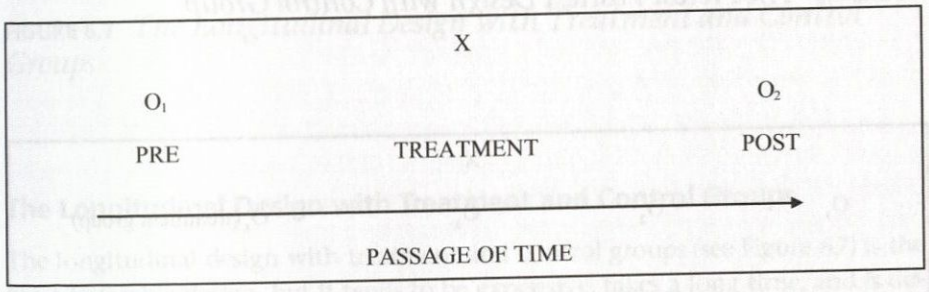

FIGURE 6.4 *The Simple Pretest-Posttest Design*

The Simple Pretest-Posttest Design

The simple pretest-posttest design (see Figure 6.4) is an easy-to-use design, but it is not a good one. Because there is no comparison or control group, we cannot adequately determine whether the program or some other unmeasured influence (confounding factor) produced the observed change from O_1 to O_2. How do we know, for example, if self-esteem wouldn't have increased (or decreased) even without the intervention? How do these clients compare to a similar group who didn't receive the intervention?

The Pretest-Posttest Design with Control Group

The pretest-posttest design with control group (see Figure 6.5) is a much better design than the simple pretest-posttest design. A control group gives us some means of comparing initial measures with later measures. If the two groups are relatively equivalent on variables likely to influence the outcome measure, we can compare the outcomes observed for the two groups to evaluate program impact. If change in the outcome measure for the treatment group is significantly better than change on the same measure for the comparison group, we have evidence that the program is effective.

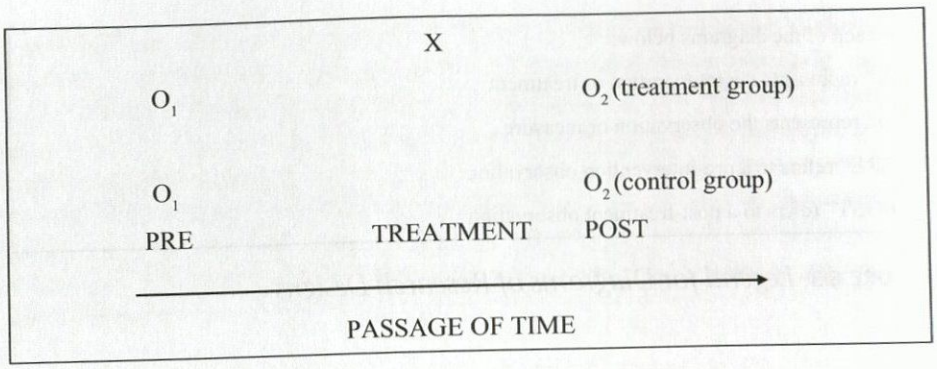

FIGURE 6.5 *The Pretest-Posttest Design with Control Group*

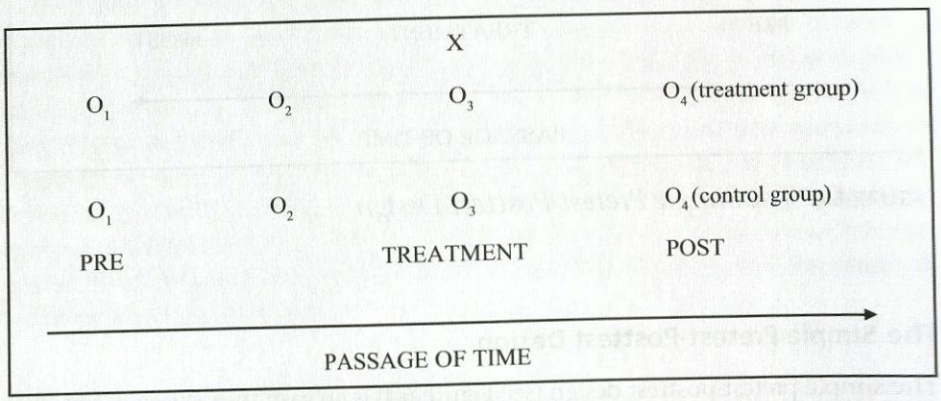

FIGURE 6.6 *The Pretest-Posttest Design with Multiple Pretests*

The Pretest-Posttest Design with Multiple Pretests

The pretest-posttest design with multiple pretests (see Figure 6.6) is a slight improvement over the pretest-posttest with control group design. Since some conditions such as attitudes are variable, it gives us a better assessment of clients' and nonclients' condition before treatment. This design allows us to obtain a baseline of behavior or attitudes for each group before the intervention begins. A baseline is always preferable to a one-shot (cross-sectional) assessment of pre-intervention characteristics. This is valuable because it gives us a much better indication of how stable or unstable the specific behavior we might be interested in is, and whether treatment or comparison groups differ in their baselines prior to treatment. In Figure 6.6, you can see that the design calls for separate but identical assessments of the outcome measure before treatment begins.

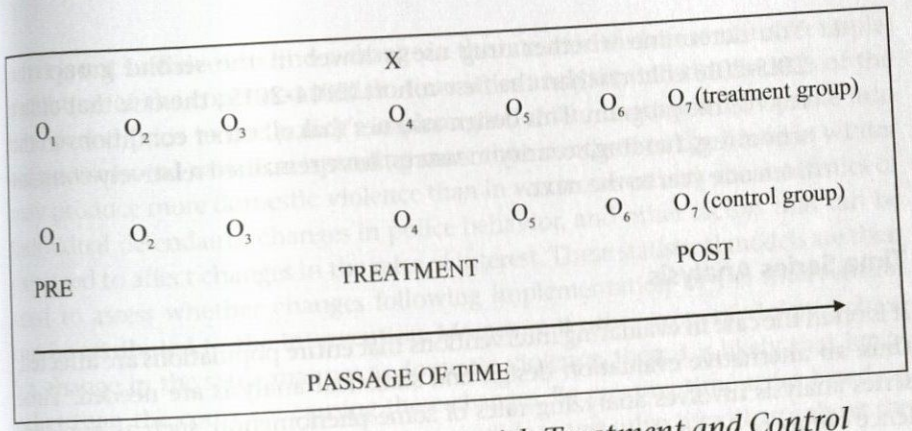

FIGURE 6.7 *The Longitudinal Design with Treatment and Control Groups*

The Longitudinal Design with Treatment and Control Groups

The longitudinal design with treatment and control groups (see Figure 6.7) is the most favorable design, but it tends to be expensive, takes a long time, and is difficult to conduct. It is the most favorable because it gives us a baseline for both the treatment and control groups, both before and after the intervention begins. With this design, we measure the outcome at several points after the treatment has ended in order to be certain that our outcome results are stable.

The Cohort Design

When it is difficult to actually assign clients to treatment and control groups, the evaluator may decide to use a cohort design (see Figure 6.8), which uses similar groups of people who go through the same system or experience but at different times. One cohort gets the intervention program; the other cohort (the control group) doesn't. This design is often used for school studies. For example, you've identified the ninth grade in one school as a group particularly vulnerable to experimentation with drugs, but still "reachable." The following steps might be taken to set up a cohort design.

1. Measure self-reported drug use of the ninth-grade class at the end of the 2014–2015 school year.
2. At the beginning of the next school year (2015–2016), you start a drug awareness education program in the new ninth grade.
3. At the end of the school year (2015–2016), you measure the level of drug use in this second cohort.

4. You determine whether drug use is lower in the second group (the 2015–2016 cohort) than the first cohort (2014–2015), the one that didn't receive the program. This design assumes that all other conditions in the school (e.g., funding, security measures) have remained relatively constant from one year to the next.

Time Series Analysis

It is often the case in evaluating interventions that entire populations are affected; thus, an alternative evaluation design and statistical analysis are needed. Time series analysis involves analyzing rates of some phenomenon, such as gun violence events or domestic violence arrests, within a population over a long period of time prior to implementation of a policy and then comparing those data to the same rate information for a period of time following implementation. The data points must be separated by precisely the same period of time, such as monthly or quarterly rates (Choate et al., 2015; Braga et al., 2008; Gover et al., 2003). In some studies, the same process is conducted with a comparison group, but often the pre-intervention period of time provides the comparison.

Typically the pre-intervention data exist prior to the study. Gover et al., 2003; see note 18), for example, found that creation of a domestic violence court in Lexington County, South Carolina, significantly reduced rates of recidivism among defendants arrested for domestic violence. In this study, they used 60 months of data to construct a time series (similar to a trend line). The domestic violence court was implemented at the end of month 34, thus interrupting the status quo. Then recidivism data were collected for the remaining 26 months. This interruption provided a picture of what effect the domestic violence court had on the typical recidivism trend.

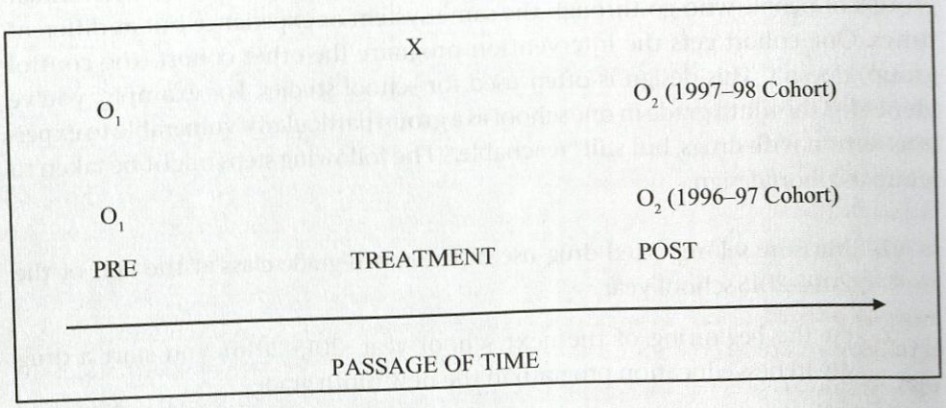

FIGURE 6.8 *The Cohort Design*

It is not sufficient to find a change in the rate months before and after implementation of the court occurred. In time series analysis, statistical models of the time period prior to the policy's implementation are constructed that take into account seasonal variation (spending more time confined to the house in winter may produce more domestic violence than in warmer weather), characteristics of individual defendants, changes in police behavior, and other factors that can be expected to affect changes in the rates of interest. These statistical models are then used to assess whether changes following implementation of the intervention can be attributed to the intervention. Moreover, if other types of violent behavior change in the same manner as domestic violence, then it is likely that forces other than the court are causing these changes. By creating time series based on these alternative forms of violence, it is possible to test other hypotheses about the reason for changes in domestic violence recidivism. The authors of this study were able to conclude that the domestic violence court reduced recidivism among these defendants.

IDENTIFY USERS AND USES OF EVALUATION RESULTS

Who will be interested in the results of this evaluation, and how will they use the information? If any evaluation is to be useful, it should serve the information needs of the program or policy and its stakeholders (see Chapter 2). Major stakeholders include the funding agency, but any intervention has multiple stakeholders, such as citizens, politicians, criminal justice officials, volunteers, targets, and so on. The time spent previously (at Stages 2 and 3) identifying and communicating with stakeholders should not be wasted. Evaluation is a critical means of demonstrating accountability, and hopefully effectiveness, to stakeholders.

Typical evaluation uses are the expansion and replication of successful programs and policies, elimination of unsuccessful ones, and an investment in developing promising programs. Evaluation results can be used solely to make judgments about programs and policies, or they can become part of a process of continuous improvement. In the latter case, changes are made to the program that are thought to improve chances of better results, and then this revised program design is retested.

The change agent should develop plans and assign responsibility for packaging and communicating evaluation results to different users. It is particularly important to find means of communication that different audiences can understand. Even at academic conferences, for instance, many eyes in the audience glaze over when a presenter puts up overheads cluttered with complicated statistical results. If the results are to be useful, and used, one must create means of communication that intended audiences could understand and react to. Stakeholders must be able to participate in a dialogue with the report writer or presenter.

DISCUSSION QUESTIONS

1. What is meant by the term "evaluation"?
2. What are the main differences between performance and impact evaluations?
3. How can an automated information system increase the availability of evaluation data?
4. Define and describe an example of efficiency analysis.
5. What are the three prerequisites for evaluation? Explain.
6. What can we learn from developing a logic model?
7. Define: (a) reliability, and (b) validity.
8. (a) Define confounding factors. (b) Describe the three most common types of confounding factors.
9. Refer back to Example 6.1 (the Minneapolis Domestic Violence Experiment). How were confounding factors illustrated by this example?
10. How does one minimize possible confounding effects in an evaluation? Make sure you discuss the two major techniques.
11. How are intermediate (short-term) and ultimate outcomes different?
12. Describe each of the following research designs: (a) simple pretest-posttest, (b) pretest-posttest with a control group, (c) pretest-posttest with multiple pretests, (d) longitudinal design with treatment and control groups, (e) cohort design, and (f) time series analysis.

ercise 6.1

ur instructor may ask you to analyze a published evaluation study. One good example is Evalua-
n of Comprehensive Services for Victims of Human Trafficking: Key Findings and Lessons Learned,
nducted by Caliber (2007), a major research center, and available at https://www.ncjrs.gov/pdf
s1/nij/grants/218777.pdf. This report is particularly valuable since it also includes an evaluability
essment. Retrieve and read the article and answer the following questions: (a) what were the
ogram's intended outcomes? (b) what confounding factors were addressed by the researchers?
what research methods were used? and (d) to what degree, if any, did the program achieve its
ended outcome objectives? Give specific examples and evidence from the article to support your
swer.

Case Study 6.1 Evaluation of an Illegal Immigration Enforcement Policy

Instructions: *Read the case study below, and then answer the questions that follow.*

Although some research designs are stronger than others, evaluation studies often make use of multiple methods in order to answer different questions or to increase the validity of conclusions drawn from the study. One such study is the evaluation of an illegal immigration enforcement policy that was implemented in Prince William County, Virginia. Prince William County is a wealthy section of Virginia and part of the Washington, DC, metropolitan area (Guterbock et al., 2010). The evaluation involved three research centers and multiple methods to address several questions regarding the impact of a new policy. Between 2000 and 2006, Prince William County experienced a large increase in the number of Hispanic residents. Concerns were raised not only about the possibility that many of these new residents were illegal immigrants, but that crimes associated with this population were also increasing. The County Board soon passed an illegal immigration law requiring police officers to inquire about the status of any person *stopped* for any reason and for whom there was probable cause to suspect their immigration status. It also reduced county services to all persons found to be illegal immigrants. After only two months, the law enforcement part of the policy was revised to target only persons *arrested* for a law violation, thus narrowing the number of potential illegal status checks. This was done to avoid claims of ethnic profiling.

Goals and Outcomes

The illegal immigration policy was designed to reduce the number of illegal immigrants, street crime, and public disorder, and to reduce chances that county revenues were being expended on services to illegal immigrants. At the same time, the County Board wanted to protect the general reputation of the county. Of course, there remained the possibilities that some subjects of the policy would protest their treatment and sue the county for violations of their civil rights and that desired relations with the Hispanic community would deteriorate.

Methods Used

This study focused primarily on the police-related provisions and procedures of the policy. Thus, the researchers measured effects on the police department and effects on the community that resulted from police activity. Obviously, there was no opportunity for use of an experimental design, and use of comparison groups was limited, but efforts were made to compare changes in several types of crime rates in Prince William County with adjacent counties, using UCR data.

The quantitative analysis depended mainly on data reported by the county police department to the FBI's Uniform Crime Reports (UCR) system, UCR data on adjacent counties, crime data from the [Washington] Metropolitan Council of Governments, a two-wave survey of all uniformed police