CHAPTER TEN

Quasi-Experimental Designs and Program Evaluation

CHAPTER OUTLINE

Overview

In the most general sense, an experiment is a test; it is a procedure we use to find out something that we don't yet know. In this sense we experiment when we add new ingredients to a chili recipe in order to see whether they improve its taste. We experiment with new ways to catch fish by changing the lures we use. We experiment when we take a different route to our job in order to find a faster way to commute. As you no doubt recognize, however, these kinds of informal "experiments" are much different from the experiments that are typically carried out in psychological research. Experimental methods, unlike other research techniques such as observation and surveys, are viewed as the most efficient way to determine causation. But determining causation is not always easy, and in the last few chapters you were introduced to the complexity of the task facing researchers who seek to understand a phenomenon by discovering what caused it.

In this chapter we continue our discussion of experimental methods, but we focus on experiments as they are conducted in natural settings such as hospitals, schools, and businesses. You will see that the task of drawing cause-effect conclusions in these settings often becomes even more difficult, and that new problems arise when an investigator leaves the confines of the laboratory to do experiments in natural settings.

There are many reasons why researchers do experiments in natural settings. One reason for these "field experiments" is to test the external validity of a laboratory finding (see Chapter 6). That is, we seek to find out if a treatment effect observed in the laboratory works in a similar way in another setting. Other reasons for experimenting in natural settings are more practical. Research in natural settings is likely to be associated with attempts to improve conditions under which people live and work. The government may experiment with a new tax system or a new method of job training for the economically disadvantaged. Schools may experiment by changing lunch programs, after-school care, or curricula. A business may experiment with new product designs, methods of delivering employee benefits, or flexible work hours. In these cases, as is true in the laboratory, it is important to determine whether the "treatment" caused a change. Did a change in the way patients are admitted to a hospital emergency room cause patients to be treated more quickly and efficiently? Did a college energy conservation program cause a decrease in energy consumption? Knowing whether a treatment was effective permits us to make important decisions about continuing the treatment, about spending additional money, about investing more time and effort, or about changing the present situation on the basis of our knowledge of the results. Research that seeks to determine the effectiveness of changes made by institutions, government agencies, and other organizations is one goal of *program evaluation.*

In this chapter we describe obstacles to doing experiments in natural settings, and we discuss ways of overcoming these obstacles so that true experiments are done whenever possible. Nevertheless, true experiments are sometimes not feasible outside the laboratory. In these cases, experimental procedures that only

approximate the conditions of laboratory experiments must be considered. We discuss several of these *quasi-experimental* techniques. We conclude by providing a brief introduction to the logic, procedures, and limitations of program evaluation.

True Experiments

Characteristics of True Experiments

- In true experiments, researchers manipulate an independent variable with treatment and comparison condition(s) and exercise a high degree of control (especially through random assignment to conditions).

As we have noted, although many everyday activities (such as altering the ingredients of a recipe) might be called experiments, we would not consider them "true" experiments in the sense in which experimentation has been discussed in this textbook. Analogously, many "social experiments" carried out by the government and those that are conducted by company officials or educational administrators are also not true experiments. *A true experiment is one that leads to an unambiguous outcome regarding what caused an event.*

True experiments exhibit three important characteristics:

1 In a true experiment some type of intervention or treatment is implemented.

2 True experiments are marked by the high degree of control that an experimenter has over the arrangement of experimental conditions, assignment of participants, systematic manipulation of independent variables, and choice of dependent variables. The ability to assign participants randomly to experimental conditions is often seen as the most critical defining characteristic of the true experiment (Judd, Smith, & Kidder, 1991).

3 Finally, true experiments are characterized by an appropriate comparison. Indeed, the experimenter exerts control over a situation to establish a proper comparison to evaluate the effectiveness of a treatment. In the simplest of experimental situations, this comparison is one between two comparable groups that are treated exactly alike except for the variable of interest.

When the conditions of a true experiment are met, any differences in a dependent variable that arise can logically be attributed to the differences between levels of the independent variable. There are differences, however, between true experiments done in natural settings and experiments done in a laboratory. A few of the most important differences are described in Box 10.1.

Obstacles to Conducting True Experiments in Natural Settings

- Researchers may experience difficulty obtaining permission to conduct true experiments in natural settings and gaining access to participants.
- Although random assignment is perceived by some as unfair because it may deprive individuals of a new treatment, it is still the best way and fairest way to determine if a new treatment is effective.

BOX 10.1

DIFFERENCES BETWEEN EXPERIMENTS IN THE LAB AND IN NATURAL SETTINGS

Experiments that are conducted outside the laboratory are likely to differ in a number of significant ways from those done in the laboratory. Not every experiment in a natural setting differs from laboratory experiments in all of these ways, of course. But if you are thinking of doing research in a natural setting, we urge you to consider the following critical issues.

Control

More than anything else, the scientist is concerned with control. Only by controlling those factors that are assumed to influence a phenomenon can we make a decision about what caused it. For instance, random assignment of participants to conditions of an experiment is a method of control used to balance individual differences across conditions. Or, researchers can hold other factors constant that are likely to influence a phenomenon. In a natural setting, a researcher may not always have the same degree of control over assignment of participants or over the conditions of an experiment that she or he would have in a laboratory. A researcher may even be asked to evaluate whether an intervention was effective without having been involved in the planning or conduct of the "experiment." This kind of "after-the-fact" evaluation is especially difficult because those conducting the study may not have considered important factors in the planning and execution of the intervention.

External validity

The high degree of control in the "artificial" environment of the laboratory that increases the internal validity of research often decreases the external validity of the findings. Experiments in natural settings may therefore need to be done in order to establish the external validity of a laboratory finding. When an experiment is done primarily to test a specific psychological theory, however, the external validity of a laboratory finding may not be all that important (e.g., Mook, 1983). In contrast, the external validity of research done in natural settings is often very important. This is especially true when social experimentation serves as the basis for large-scale social changes, such as trying out new ways to curb drunk driving or new procedures for registering voters. Will the results of a program that is judged to be beneficial for curbing drunk driving in a midwestern state generalize to states in other areas of the country? These are, of course, questions about the external validity of research findings.

Goals

Experimentation in natural settings often has different goals from those of laboratory research (see Chapter 2). Laboratory research frequently represents *basic research* with the single goal of understanding a phenomenon—of determining how "nature" works. It may be done to gain knowledge merely for knowledge's sake. *Applied research* is also directed toward discovering the reasons for a phenomenon, but it is likely to be done only when knowing the reasons for an event will lead to changes that will improve the present situation. Experimentation in natural settings, therefore, is more likely than laboratory research to have practical goals.

Consequences

Sometimes experiments are conducted that have far-reaching impact on communities and society, affecting large numbers of people. The Head Start program for disadvantaged children and the *Sesame Street* television show were social experiments designed to improve the education of hundreds of thousands of children across the nation (see Figure 10.1). Social experiments are also carried out on a smaller scale in natural settings such as in local schools or businesses. Clearly, society's "experiments" are likely to have consequences of greater immediate impact than those of laboratory research. By contrast, the immediate consequences of a laboratory experiment can be substantial, but they are much more likely to be minimal. They may directly affect only the lives of a few researchers and of those relatively few participants recruited to participate.

FIGURE 10.1 As a social experiment, *Sesame Street* was designed to improve the education of hundreds of thousands of children.

Experimental research is an effective tool for solving problems and answering practical questions. Nevertheless, two major obstacles often arise when we try to carry out experiments in natural settings. The first problem is obtaining permission to do the research from individuals in positions of authority. Unless they believe that the research will be useful, school board presidents and government and business leaders are unlikely to support research financially or otherwise. The second, and often more pressing, obstacle to doing experiments in natural settings is the problem of access to participants. This problem can prove especially troublesome if participants are to be randomly assigned to either a treatment group or a comparison group.

Random assignment to conditions appears unfair at first—after all, random assignment requires that a potentially beneficial treatment be withheld from some participants. Suppose that a new approach to the teaching of foreign languages was to be tested at your college or university. Suppose further that, when you went to register for your next semester's classes, you were told that you would be randomly assigned to one of two sections taught at the time you selected—one section involving the old method and one involving the new method. How would you react? Your knowledge of research methods tells you that the two methods must be administered to comparable groups of students and that random assignment is the best way to ensure

such comparability. Nonetheless, you might be tempted to feel that random assignment is not fair, especially if you are assigned to the section using the old (old-fashioned?) method. Let's take a closer look at the fairness of random assignment.

If those responsible for selecting the method of foreign language instruction already knew that the new method was more effective than the old method at schools such as yours, there would be little justification for testing the method again. Under such circumstances we would agree that withholding the new method from students in the control group would be unjust. If we do not know whether the new method is better, however, any approach other than conducting a true experiment will leave us in doubt about the new method's effectiveness. Random assignment to treatments—call it a "lottery" if you prefer—may be the fairest procedure for assigning students to sections. The old method of instruction, after all, was considered effective before the development of the new method. If the new method proves less effective, random assignment will have actually "protected" the control participants from receiving an ineffective treatment.

There are ways to offer a potentially effective treatment to all participants while still maintaining comparable groups. One way is to alternate treatments. For example, Atkinson (1968) randomly assigned students to receive computer-assisted instruction (the treatment) in either English or math and then tested both groups in English and math. Each group served as a control for the other on the test for which its members had not received computer-assisted instruction. After completing the experiment, both groups could then be given computer-assisted instruction in the subject matter to which they had not been previously exposed. Thus, all participants received all potentially beneficial treatments.

Establishing a proper control group is also possible if there is more demand for a service than an agency can meet. People who are waiting to receive the service can become a *waiting-list control group.* It is essential, however, that people be assigned to the waiting list randomly. People who are first in line are no doubt different on important dimensions from those who arrive last (e.g., more eager for treatment). Random assignment is necessary to distribute these characteristics in an unbiased way between treatment and comparison groups.

There will always be circumstances in which random assignment simply cannot be used. For example, in clinical trials involving tests of new medical treatments, it may be extremely difficult to get patients to agree to be randomly assigned to either the treatment group or the control (no treatment) group. As you will see, *quasi-experimental designs* can be used in these situations. The logic and procedures for these quasi-experimental designs will be described later in this chapter.

Threats to Internal Validity Controlled by True Experiments

- Threats to internal validity are confounds that serve as plausible alternative explanations for a research finding.
- Major classes of threats to internal validity include history, maturation, testing, instrumentation, regression, subject attrition, selection, and additive effects with selection.

Prior to doing an experiment, we want to consider what major classes of possible explanations can be ruled out by our experimental procedure. Only by controlling all possible alternative explanations can we arrive at a definite causal inference. In previous chapters, we referred to various uncontrolled factors that threaten the internal validity of an experiment as confounding factors (they are also called confounds). Several types of confounds were identified in earlier chapters (see especially Chapter 6). Campbell and Stanley (1966; Cook & Campbell, 1979; see also Shadish, Cook, & Campbell, 2002; West, 2010) have identified eight classes of confounds that they call **threats to internal validity.** You have already been introduced to some of these; others will be new. After reviewing these major threats to internal validity, we will be able to judge the extent to which various experimental procedures control for these kinds of alternative explanations of a treatment effect.

Key Concept

History The occurrence of an event other than the treatment can threaten internal validity if it produces changes in the research participants' behavior. A true experiment requires that participants in the experimental group and in the control group be treated the same (have the same history of experiences while in the experiment) except for the treatment. In the laboratory, this is usually accomplished by balancing or holding conditions constant. When doing experiments in natural settings, however, the researcher may not be able to maintain a high degree of control, so confounding due to history can threaten internal validity. For example, suppose that you set out to test whether a college-level critical thinking course does, in fact, change students' thinking. And suppose further that you simply examined students' performance on a critical thinking test at the beginning of the course and then again at the end of the course. Without an appropriate comparison group, **history** would be a threat to internal validity if events other than the treatment (i.e., the critical thinking course) occurred that might improve students' critical thinking abilities. For instance, suppose many students in the course also accessed a website designed to teach critical thinking that wasn't required for the course. The students' history, now including the website experience, would confound the treatment and therefore pose a threat to the internal validity of the study.

Key Concept

Maturation Participants in an experiment necessarily change as a function of time. They grow older, become more experienced, and so forth. Change associated with the passage of time per se is called maturation. For example, suppose a researcher is interested in evaluating children's learning over a school year using a new teaching technique. Without a proper comparison, a researcher might attribute the changes in children's performance between the beginning and the end of the school year to the effect of the teaching intervention when, in reality, the changes were simply due to a **maturation** threat to validity. That is, the children's learning may have improved simply because their cognitive abilities increased as they aged.

Key Concept

Testing Taking a test generally has an effect on subsequent testing. Consider, for example, the fact that many students often improve from the initial test in a

course to the second test. During the first test the students gain familiarity with the testing procedure and with the instructor's expectations. This familiarity then affects their performance on the second test. Likewise, in the context of a psychology experiment in which more than one test is given (e.g., in a pretest-posttest design), **testing** is a threat to internal validity if the effect of a treatment cannot be separated from the effect of testing.

Key Concept

Instrumentation Changes over time can take place not only in the participants of an experiment (e.g., maturation or increased familiarity with testing), but also in the instruments used to measure participants' performance. This is most clearly a possibility when human observers are used to assess behavior. For instance, observer bias can result from fatigue, expectations, and other characteristics of observers. Unless controlled for, these changes in the observers represent an **instrumentation** threat to internal validity by providing alternative explanations for differences in behavior between one observation period and another. Mechanical instruments also may change with repeated use. A researcher known to the authors once found that a machine used to present material in a learning experiment was not working the same at the end of the experiment as it was at the beginning. Measures made near the end of the experiment differed from those made at the beginning of the experiment. Thus, what looked like a learning effect was really just a change in the instrument used to measure learning.

Key Concept

Regression Statistical **regression** is always a problem when individuals have been selected to participate in an experiment because of their "extreme" scores. Extreme scores on one test are not likely to be as extreme on a second test. In other words, a very, very bad performance, or a very, very good performance (both of which we have all experienced), is likely to be followed by a performance that is not quite so bad, or not quite so good, respectively. Consider, for instance, your best ever performance on a classroom examination. What did it take to "nail" this test? It took, no doubt, a lot of hard work. But it is also likely that some luck was involved. Everything has to work just right to produce an extremely good performance. If we are talking about an exam, then it is likely that the material tested was that which you just happened to study the hardest, or the test format was one you particularly like, or it came at a time when you were feeling particularly confident, or all of these and more. Particularly good performances are "extreme" because they are inflated (over our usual or typical performance) by chance. Similarly, an especially bad test performance is likely to have occurred because of some bad luck. When tested again (following either a very good or a very bad performance), it is simply not likely that chance factors will "gang up" the same way to give us that super score or that very poor score. We will likely see a performance closer to the average of our overall scores. This phenomenon frequently is called *regression to the mean*. Statistical regression is more likely when a test or measure is unreliable. When an unreliable test is used, we can expect scores to be inconsistent over time.

Key Concept

Now, consider an attempt to raise the academic performance of a group of college students who performed very poorly during their first semester of college (the "pretest"). Participants are selected because of their extreme performance

(in this case, extremely poor performance). Let us assume that a treatment (e.g., a 10-hour study skills workshop) is then applied. Statistical regression is a threat to internal validity because we would expect these students to perform slightly better after the second semester (the "posttest") *without any treatment* simply due to statistical regression. An unknowing researcher may mistakenly confuse this "regression effect" with a "treatment effect."

Key Concept

Subject Attrition As discussed in Chapter 6, a threat to internal validity occurs when participants are lost from an experiment, for example, when participants drop out of the research project. The **subject attrition** threat to internal validity rests on the assumption that the loss of participants changes the nature of the group that was established prior to the treatment—for example, by destroying the equivalence of groups established through random assignment. This might occur, for instance, if an experimental task is very difficult and causes some experimental participants to become frustrated and to drop out of the experiment. Participants who are left in the experimental group will differ from those who dropped out (and possibly from those in a control group) if for no other reason than that they were able to do the task (or at least stuck it out).

Key Concept

Selection When, from the outset of a study, differences exist between the kinds of individuals in one group and those in another group in the experiment, there is a threat to internal validity due to **selection.** That is, the people who are in the treatment group may differ from people in the comparison group in many ways other than their group assignment. In the laboratory, this threat to internal validity is generally handled by balancing participant characteristics through random assignment. When one is doing experiments in natural settings, there are often many obstacles to randomly assigning participants to treatment and comparison conditions. These obstacles prevent doing a true experiment and hence present a possible threat to internal validity due to selection.

Additive Effects with Selection Individual threats to internal validity such as history and maturation can be a source of additional concern because they can combine with the selection threat to internal validity. Specifically, when comparable groups are not formed by random assignment, there are possible problems due to additive effects of (1) selection and maturation, (2) selection and history, and (3) selection and instrumentation. For example, *additive effects of selection and maturation* could occur if first-year students in college who served as an experimental group were compared with sophomores who served as a control group. Changes in students that occur during their first year (as students gain familiarity with the college environment) might be presumed to be greater than the changes that occur during the sophomore year. These differences in maturation rates might explain any observed differences between the experimental and control groups, rather than the differences being due to the experimental intervention.

An *additive effect of selection and history* results when events occurring in time have a different effect on one group of participants than on another. This is particularly a problem when intact groups are compared. Perhaps due to events

that are peculiar to one group's situation, an event may have more of an impact on that group than on another. Consider, for example, research involving an investigation of the effectiveness of an AIDS awareness campaign involving two college campuses (one treatment and one control). Nationwide media attention to AIDS might reasonably be assumed to affect students on both campuses equally. However, if a student with AIDS died at one college during the study and the story was featured in the college newspaper, we would assume that research participants at this student's college would be affected differently compared to those at the other. In terms of assessing the effect of an AIDS awareness campaign, this situation would represent an additive effect of selection and history.

Finally, an *additive effect of selection and instrumentation* might occur if a test instrument is relatively more sensitive to changes in one group's performance than to changes in another's. This occurs, for instance, when ceiling or floor effects are present. Such is the case when a group scores initially so low on an instrument (floor effect), that any further drop in scores cannot be reliably measured, or so high (ceiling effect) that any more gain cannot be assessed. As you can imagine, a threat to internal validity would be present if an experimental group showed relatively no change (due to floor or ceiling effects), while a control group changed reliably because its mean performance was initially near the middle of the measurement scale.

One of the great advantages of true experiments is that they *control* for all these threats to internal validity. As Campbell (1969) emphasizes, true experiments should be conducted when possible, but if they are not feasible, quasi-experiments should be conducted. "We must do the best we can with what is available to us" (p. 411). Quasi-experiments represent the best available compromise between the general aim of gaining valid knowledge regarding the effectiveness of a treatment and the realization that true experiments are not always possible.

Problems That Even True Experiments May Not Control

- Threats to internal validity that can occur in any study include contamination, experimenter expectancy effects, and novelty effects.
- Contamination occurs when information about the experiment is communicated between groups of participants, which may lead to resentment, rivalry, or diffusion of treatment.
- Novelty effects occur when people's behavior changes simply because an innovation (e.g., a treatment) produces excitement, energy, and enthusiasm.
- Threats to external validity occur when treatment effects may not be generalized beyond the particular people, setting, treatment, and outcome of the experiment.

Before considering specific quasi-experimental procedures, we should point out that even true experiments may not control for all possible threats to the interpretation of an experimental outcome. Although major threats to internal validity are eliminated by the true experiment, there are some additional threats that the investigator who is working in natural settings must guard against. We

BOX 10.2

EXPERIMENTAL CONTAMINATION

There are several possible effects resulting from communication between groups of experimental participants. These include (1) *resentment* on the part of individuals receiving less desirable treatments, (2) *rivalry* among groups receiving different treatments, and (3) a general *diffusion of treatments* across the groups (see Cook & Campbell, 1979; Shadish et al., 2002).

- *Resentment* Consider a situation in which individuals have been randomly assigned to a control group. Further, assume that control group participants learn that "other" participants are receiving a beneficial treatment. What do you think might be the reaction of the control participants? One possibility is that the control participants will feel resentful and demoralized. As Cook and Campbell explain, in an industrial setting the person receiving the less desirable treatment may retaliate by lowering productivity. In an educational setting, teachers or students might "lose heart" or become angry. This effect of "leaked" information about a treatment may make a treatment look better than it ordinarily would because of the lowered performance of the control group that responds with resentment.
- *Rivalry* Another possible effect that may occur when a control group learns about another group's good fortune is a spirit of competition or rivalry. That is, a control group might become motivated to reduce the expected difference between itself and the treatment group. As Cook and Campbell point out, this may be likely when intact groups (such as departments, work crews, branch offices, and the like) are assigned to various conditions. Realizing that another group will look better depending on how much it distinguishes itself from the control group, participants comprising the control group may be motivated to "try harder" so as not to look bad by comparison.
- *Diffusion of treatments* Yet another possible effect of contamination is diffusion of treatments. According to Cook and Campbell, this occurs when participants in a control group use information given to others to help them change their own behavior. For example, control participants may use the information given to participants in the treatment group to imitate the behavior of individuals who were given the treatment. Of course, this reduces the differences between the treated and untreated groups and affects the internal validity of the experiment.

Key Concept

will use the term *contamination* to describe one general class of threats to internal validity. **Contamination** occurs when there is communication of information about the experiment between groups of participants. Box 10.2 describes the several unwanted effects that can occur with contamination.

True experiments can also be affected by threats due to *experimenter expectancy effects* that occur when an experimenter unintentionally influences the results. Observer bias occurs when researchers' biases and expectancies lead to systematic errors in observing, identifying, recording, and interpreting behavior. (Various ways to control observer or experimenter effects were outlined in Chapter 4 and Chapter 6, e.g., using a double-blind procedure.)

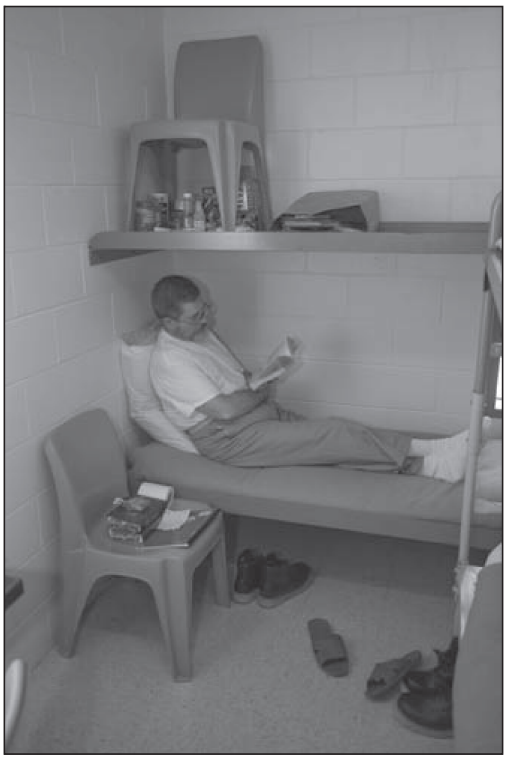
Key Concept

Novelty effects can occur when an innovation, such as an experimental treatment, is introduced (Shadish et al., 2002). For example, if little in the way of change or innovation has occurred for some time at a work site, employees may become excited or energized by the novelty (or newness) of their work environment when an intervention is introduced. Employees' newfound enthusiasm, rather than the intervention itself, may account for the "success" of the intervention. The opposite of a novelty effect can occur as a *disruption effect*, in which an innovation, perhaps with new work procedures, disrupts employees' work to such an extent that they cannot maintain their typical effectiveness.

One specific novelty effect has been labeled the *Hawthorne effect*. This refers to changes in people's behavior brought about by the interest that "significant others" show in them. The effect was named after events occurring at the Hawthorne plant of the Western Electric Company in Cicero, Illinois, near Chicago, between 1924 and 1932 (Roethlisberger, 1977). Studies were conducted to examine the relationship between productivity and conditions of the workplace. In one experiment, the amount of lighting in the plant was varied and worker performance was examined. Results revealed that *both* experimental and control groups increased their productivity during the study. Although there is some controversy surrounding the exact factors responsible for this effect (e.g., Parsons, 1974), a Hawthorne effect generally refers to a change in behavior that results from participants' awareness that someone is interested in them.

As one example of the Hawthorne effect, consider a study in which prisoners are chosen to participate in research examining the relationship between changes in prison-cell conditions and attitudes toward prison life (see Figure 10.2). If positive changes in prisoners' attitudes are obtained, the results could be due to the actual changes in cell conditions that were made, or they could be due to an increase in morale because prisoners saw the prison administration as expressing concern for them. Researchers working in natural settings must be conscious of the fact that changes in participants' behavior may be partially due to their awareness that others are interested in them. Thus, you can see that a Hawthorne effect represents a specific kind of reactivity (i.e., an awareness that one is being observed), which we discussed in previous chapters (especially Chapter 4).

FIGURE 10.2 Research investigating methods for improving prison life may be subject to Hawthorne effects.



In addition to problems resulting from threats to internal validity, true experiments can be weakened by *threats to external validity*. External validity depends mainly on how representative our sample is of the persons, settings, and times to which we want to generalize. Representativeness is normally achieved through random sampling. Because random sampling is used so infrequently (see Shadish et al., 2002), however, we can rarely say that our sample of participants, or the situation in which we are making observations, or the times during which we test individuals are representative samples of all persons, settings, treatments, or outcomes. Therefore, the investigator must be aware of possible interactions between the independent variable of an experiment and, for example, the type of individual or the nature of the setting that is involved in the experiment. Is a difference, for instance, between an experimental group and a control group that is observed with volunteers from an inner-city school in the winter also likely to be found when nonvolunteers are tested in a suburban school in the spring of the year?

Cook and Campbell describe several approaches to evaluating threats to external validity; the most important is attempting to determine the representativeness of the sample. They point out, however, that *the best test of external validity is replication*. Thus, the question of external validity is best answered by repeating the experiment with different types of participants, in different settings, with different treatments, and at different times. Occasionally partial replications can be "built into" an experiment—for example, by selecting more than one group to participate. Testing schoolchildren from a lower socioeconomic group and a higher socioeconomic group in an experiment designed to determine the effectiveness of a new educational program would provide evidence of the generality of the treatment's effectiveness across these two socioeconomic groups.

Quasi-Experiments

- Quasi-experiments provide an important alternative when true experiments are not possible.
- Quasi-experiments lack the degree of control found in true experiments; most notably, quasi-experiments typically lack random assignment.
- Researchers must seek additional evidence to eliminate threats to internal validity when they do quasi-experiments rather than true experiments.
- The one-group pretest-posttest design is called a pre-experimental design or a bad experiment because it has so little internal validity.

Key Concept

A dictionary will tell you that one definition of the prefix *quasi-* is "resembling." Quasi-experiments involve procedures that *resemble* those of true experiments. Generally speaking, **quasi-experiments** include some type of intervention or treatment and they provide a comparison, but they lack the degree of control found in true experiments. Just as randomization is the hallmark of true experiments, so *lack of randomization* is the hallmark of quasi-experiments. As Campbell and Stanley (1966) explain, quasi-experiments arise when researchers lack the control necessary to perform a true experiment.

Quasi-experiments are recommended when true experiments are not feasible. Some knowledge about the effectiveness of a treatment is more desirable than none. The list of possible threats to internal validity that we reviewed earlier can be used as a checklist in deciding just how good that knowledge is. Moreover, the investigator must be prepared to look for additional kinds of evidence that might rule out a threat to internal validity that is not specifically controlled in a quasi-experiment. For example, suppose that a quasi-experiment does not control for history threats that would be eliminated by a true experiment. The investigator may be able to show that the history threat is implausible based on a logical analysis of the situation or based on evidence provided by a supplementary analysis. If the investigator can show that the history threat is implausible, then a stronger argument can be made for the internal validity of the quasi-experiment. Researchers must recognize the specific shortcomings of quasi-experimental procedures, and they must work like detectives to provide whatever evidence they can to overcome these shortcomings. As we begin to consider the appropriate uses of quasi-experiments, we need to acknowledge that there is a great difference between the power of the true experiment and that of the quasi-experiment. *Before facing the problems of interpretation that result from quasi-experimental procedures, the researcher should make every effort possible to approximate the conditions of a true experiment*.

Perhaps the most serious limitation researchers face in doing experiments in natural settings is that they are frequently unable to assign participants randomly to conditions. This occurs, for instance, when an intact group is singled out for treatment and when administrative decisions or practical considerations prevent randomly assigning participants. For example, children in one classroom or school and workers at a particular plant represent intact groups that might receive a treatment or intervention without the possibility of randomly assigning individuals to conditions. If we assume that behavior of a group is measured both before and after treatment, such an "experiment" can be described as follows:

$$O_1 \quad X \quad O_2$$

where O_1 refers to the first observation of a group, or pretest, X indicates a treatment, and O_2 refers to the second observation, or posttest.

This *one-group pretest-posttest* design represents a pre-experimental design or, more simply, may be called a bad experiment. Any obtained difference between the pretest and posttest scores could be due to the treatment *or* to any of several threats to internal validity, including history, maturation, testing, and instrumentation threats (as well as experimenter expectancy effects and novelty effects). The results of a bad experiment are inconclusive with respect to the effectiveness of a treatment. Fortunately, there are quasi-experiments that improve upon this pre-experimental design.

The Nonequivalent Control Group Design

- In the nonequivalent control group design, a treatment group and a comparison group are compared using pretest and posttest measures.

STRETCHING EXERCISE

In this exercise we ask you to consider possible threats to internal validity in this brief description of a one-group pretest-posttest design.

A psychologist interested in the effect of a new therapy for depression recruited a sample of 20 individuals who sought relief from their depression. At the beginning of the study he asked all participants to complete a questionnaire about their symptoms of depression. The mean depression score for the sample was 42.0 (the highest possible score is 63.0), indicating severe depressive symptoms. (Individuals who are not depressed typically score in the 0 to 10 range on this measure.) During the next 16 weeks the psychologist treated participants in the study with the new treatment. At the end of the treatment the participants completed the depression questionnaire again. The mean score for the posttest was 12.0, indicating that, on average, participants' depression symptoms were dramatically reduced and indicated only mild depression. The psychologist concluded that the treatment was effective; that is, the treatment caused their depressive symptoms to improve.

Cause-and-effect statements, such as the one made by this psychologist, are essentially impossible to make when the one-group pretest-posttest design is used. To understand why this is true, we ask you to think of potential threats to internal validity in this study.

1 How might a *history* effect threaten the internal validity of this study?
2 Explain how *maturation* likely plays a role in this study.
3 Are *testing* and *instrumentation* threats likely in this study?
4 Explain how *statistical regression* might influence the interpretation of these findings.

- If the two groups are similar in their pretest scores prior to treatment but differ in their posttest scores following treatment, researchers can more confidently make a claim about the effect of treatment.
- Threats to internal validity due to history, maturation, testing, instrumentation, and regression can be controlled in a nonequivalent control group design.

Key Concept

The one-group pretest-posttest design can be modified to create a quasi-experimental design with greatly superior internal validity if two conditions are met: (1) there exists a group "like" the treatment group that can serve as a comparison group, and (2) there is an opportunity to obtain pretest and posttest measures from individuals in both the treatment and the comparison groups. Campbell and Stanley (1966) call a quasi-experimental procedure that meets these two conditions a **nonequivalent control group design.** Because a comparison group is selected on bases other than random assignment, we cannot assume that individuals in the treatment and control groups are equivalent on all important characteristics (i.e., a selection threat arises). Therefore, it is essential that a pretest be given to both groups to assess their similarity on the dependent measure. A nonequivalent control group design can be outlined as follows:

$$O_1 \; X \; O_2$$
$$- - - - - -$$
$$O_1 \quad O_2$$

The dashed line indicates that the treatment and comparison groups were not formed by assigning participants randomly to conditions.

By adding a comparison group, researchers can control threats to internal validity due to history, maturation, testing, instrumentation, and regression. A brief review of the logic of experimental design will help show why this occurs. We wish to begin an experiment with two similar groups; then one group receives the treatment and the other does not. If the two groups' posttest scores differ following treatment, we first must rule out alternative explanations before we can claim that treatment caused the difference. If the groups are truly comparable, and both groups have similar experiences (except for the treatment), then we can assume that history, maturation, testing, instrumentation, and regression effects occur to both groups equally. Thus, we may assume that both groups change naturally at the same rate (maturation), experience the same effect of multiple testing, or are exposed to the same external events (history). If these effects are experienced in the *same* way by both groups, they cannot possibly be used to account for group *differences* on posttest measures. Therefore, they no longer are threats to internal validity. Thus, researchers gain a tremendous advantage in their ability to make causal claims simply by adding a comparison group. These causal claims, however, depend critically on forming comparable groups at the start of the study, and ensuring that the groups then have comparable experiences, except for the treatment. Because this is difficult to realize in practice, as we'll see, threats to internal validity due to additive effects with selection typically are not eliminated in this design.

As you approach the end of a course on research methods in psychology, you might appreciate learning about the results of a nonequivalent control group design that examined the effect of taking a research methods course on reasoning about real-life events (VanderStoep & Shaughnessy, 1997). Students enrolled in two sections of a research methods course (and who happened to be using an edition of this textbook) were compared with students in two sections of a developmental psychology course on their performance on a test emphasizing methodological reasoning about everyday events. Students in both kinds of classes were administered tests at the beginning and at the end of the semester. Results revealed that research methods students showed greater improvement than did students in the control group. Taking a research methods course improved students' ability to think critically about real-life events.

With that bit of encouraging news in mind, let us now examine in detail another study using a nonequivalent control group design. This will give us the opportunity to review both the specific strengths and limitations of this quasi-experimental procedure.

Nonequivalent Control Group Design: The Langer and Rodin Study

- Quasi-experiments often assess the overall effectiveness of a treatment that has many components; follow-up research may then determine which components are critical for achieving the treatment effect.

Langer and Rodin (1976) hypothesized that environmental changes associated with old age contribute, in part, to feelings of loss, inadequacy, and low self-esteem among the elderly. Of particular importance is the change that occurs when elderly persons move into a nursing home. Although they usually care for the elderly quite adequately in physical terms, nursing homes often provide what Langer and Rodin call a "virtually decision-free" environment. The elderly are no longer called on to make even the simplest decisions, such as what time to get up, whom to visit, what movie to watch, and the like. In a nursing home, many or most of these everyday decisions are made for the elderly, leaving them with little personal responsibility and choice.

To test the hypothesis that the lack of opportunity to make personal decisions contributes to the psychological and even the physical debilitation sometimes seen in the elderly, Langer and Rodin carried out a quasi-experiment in a Connecticut nursing home. The independent variable was the type of responsibility given to two groups of nursing home residents. One group was informed of the many decisions they needed to make regarding how their rooms were arranged, visiting, care of plants, movie selection, and so forth. These residents were also given a small plant as a gift (if they decided to accept it) and told to take care of it as they wished. This was the responsibility-induced condition. The second group of residents, the comparison group, was also called together for a meeting, but instructions for this group stressed the staff's responsibility for them. These residents also received a plant as a gift (whether they chose to have one or not) and were told the nurses would water and care for the plants for them.

Residents of the nursing home had been assigned to a particular floor and room on the basis of availability, and some residents had been there for a long time. As a consequence, randomly assigning residents to the two responsibility groups was impractical—and probably undesirable from the administration's point of view. Therefore, the two sets of responsibility instructions were given to residents on two different floors of the nursing home. These floors were chosen, in the words of the authors, "because of similarity in the residents' physical and psychological health and prior socioeconomic status, as determined from evaluations made by the home's director, head nurses, and social worker" (Langer & Rodin, 1976, p. 193). The floors were randomly assigned to one of the two treatments. In addition, questionnaires were given to residents 1 week before and 3 weeks after the responsibility instructions. The questionnaires contained items that related to "how much control they felt over general events in their lives and how happy and active they felt" (p. 194). Furthermore, staff members on each floor were asked to rate the residents, before and after the experimental manipulation, on such traits as alertness, sociability, and activity. The investigators also included a clever posttest measure of social interest by holding a competition that asked participants to guess the number of jelly beans in a large jar. Residents entered the contest if they wished by simply filling out a piece of paper giving their estimate and name. Thus, there were a number of dependent variables to assess the residents' perceptions of control, happiness, activity, interest level, and so forth.

The Langer and Rodin study nicely illustrates the procedures of a nonequivalent control group design (see Figure 10.3). Moreover, differences between

FIGURE 10.3 Langer and Rodin (1976) used a nonequivalent control group design to study the effect of two different types of responsibility instructions on the behavior of nursing home residents. Because a "true experiment" was not conducted, the researchers examined features of the study to determine if any threats to internal validity were present.

pretest and posttest measures showed that the residents in the responsibility-induced group were generally happier, more active, and more alert following the treatment than were residents in the comparison group. Behavioral measures such as frequency of movie attendance also favored the responsibility-induced group, and, although 10 residents from this group entered the jelly bean contest, only 1 resident from the comparison group participated! The investigators point to possible practical implications of these findings. Specifically, they suggest that some of the negative consequences of aging can be reduced or reversed by giving the elderly the opportunity to make personal decisions and to feel competent.

Before turning to the specific limitations associated with this design, let us call your attention to another feature of the Langer and Rodin study, one that characterizes many experiments in natural settings. The treatment in the Langer and Rodin study actually had several components. For example, residents in the treatment group were encouraged by the staff to make decisions about a number of different things (e.g., movies, rooms, etc.), and they were offered a plant to take care of. The experiment evaluated, however, the treatment "package." That is, the effectiveness of the overall treatment, not individual components

of the treatment, was assessed. We only know (or at least we assume based on the evidence) that the treatment with all its components worked; we don't necessarily know whether the treatment would work with fewer components or whether one component is more critical than others.

Research in natural settings is often characterized by treatments with many components. Moreover, the initial goal of such research is often to assess the overall effect of the treatment "package." Finding evidence for an overall treatment effect, therefore, may be only the first stage in a research program if we want to identify the critical elements of a treatment. There may be practical as well as theoretical benefits to such identification. On practical grounds, should research reveal that only some of the treatment's features are critical to produce the effect, then perhaps the less critical features could be dropped. This may make the treatment more cost-effective and more likely to be adopted and carried out. From a theoretical standpoint, it is important to determine whether components of the treatment specified by a theory as being critical are, indeed, the critical components. When you hear about research showing an overall treatment effect you might think about how additional research could reveal what specific components are critical to the treatment's effect.

Sources of Invalidity in the Nonequivalent Control Group Design

- To interpret the findings in quasi-experimental designs, researchers examine the study to determine if any threats to internal validity are present.
- The threats to internal validity that must be considered when using the nonequivalent control group design include additive effects with selection, differential regression, observer bias, contamination, and novelty effects.
- Although groups may be comparable on a pretest measure, this does not ensure that the groups are comparable in all possible ways that are relevant to the outcome of the study.

According to Cook and Campbell (1979), the nonequivalent control group design generally controls for all major classes of potential threats to internal validity except those due to additive effects of (1) selection and maturation, (2) selection and history, (3) selection and instrumentation, and (4) those due to differential statistical regression. We will explore how each of these potential sources of invalidity might pose problems for Langer and Rodin's interpretation of their findings. We will then explain how Langer and Rodin offered both logical argument and empirical evidence to refute the possible threats to the internal validity of their study. We will also examine how experimenter bias and problems of contamination were controlled. Finally, we will comment briefly on challenges of establishing external validity that are inherent in the nonequivalent control group design.

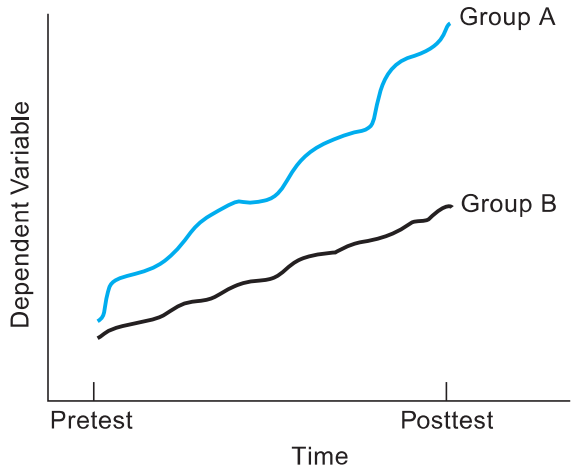
An important initial finding in Langer and Rodin's study was that the residents in the two groups did not differ significantly on the pretest measures. It would not have been surprising to find a difference between the two groups before the treatment was introduced because the residents were not randomly

assigned to conditions. Even when pretest scores show no difference between groups, however, we cannot assume that the groups are "equivalent" (Campbell & Stanley, 1966). We will explain why we cannot conclude that the groups are equivalent in the discussion that follows.

Selection-Maturation Effect An additive effect of selection and maturation occurs when individuals in one group grow more experienced, more tired, or more bored at a faster rate than individuals in another group (Shadish et al., 2002). A selection-maturation effect is more likely to be a threat to internal validity when the treatment group is self-selected (the members deliberately sought out exposure to the treatment) and when the comparison group is from a different population from the treatment group (Campbell & Stanley, 1966). Langer and Rodin selected their groups (but not individuals) randomly from the same population of individuals. Consequently, their design more closely approaches a true experiment than it would if individuals in the two groups had come from different populations (Campbell & Stanley, 1966). A selection-maturation effect would have been more likely, for example, if residents in a nursing home were compared with those attending a sheltered workshop program for the elderly, or if residents on different floors of a nursing facility require different levels of care.

The possibility of a selection-maturation effect is one reason we cannot conclude the groups are equivalent (comparable) even when pretest scores are the same on average for the treatment and control groups. The natural growth rate of two groups from different populations might be different, but the pretest may have been taken at a time when both groups happened to be about the same. This problem is illustrated in Figure 10.4. The normal rate of change is greater in Group A than in Group B, but the pretest is likely to show that the groups do not differ. Because of the differential growth rate, however, the groups would probably show a difference at the posttest that could be mistaken for a treatment effect. There is a second, and more general,

FIGURE 10.4 Possible differential growth rates for two groups (A and B) in the absence of treatment.



reason why we cannot conclude that groups are comparable based only on the absence of a difference between the groups on the pretest. The pretest is likely to measure respondents on only one measure, or at best on a few measures. The mere fact that individuals do not differ on one measure does not mean they don't differ on other measures that are relevant to their behavior in this situation.

Is there any reason to suspect a selection-maturation effect in the Langer and Rodin study? That is, would it be reasonable to expect that residents on the treatment floor would change naturally at a faster rate than would patients on the no-treatment floor? Several kinds of evidence suggest that this would not be the case. First, the procedure the nursing home used to assign residents to the two floors was basically random, and the floors were assigned randomly to the treatment and no-treatment conditions. Langer and Rodin also reported that the residents of the two floors were, on the average, equivalent on measures such as socioeconomic status and length of time at the nursing home. Finally, although it is not sufficient evidence in itself, residents on the two floors did not differ on the pretest measures. Thus, the evidence strongly indicates that there was not a threat to the internal validity of the Langer and Rodin study due to the additive effects of selection and maturation.

Selection-History Effect Another threat to internal validity that is not controlled in the nonequivalent control group design is the additive effect of selection and history. Cook and Campbell (1979) refer to this problem as *local history effects*. This problem arises when an event other than the treatment affects one group and not the other. Local history, for example, could be a problem in the Langer and Rodin study if an event affecting the residents' happiness and alertness occurred on one floor of the nursing home but not on the other. You can probably imagine a number of possibilities. A change in nursing staff on one floor, for instance, might bring about either an increase or a decrease in residents' morale, depending on the nature of the change and any differences between the behavior of a new nurse and that of the previous one. Problems of local history become more problematic the more the settings of the individuals in the treatment and comparison groups differ. Langer and Rodin do not specifically address the problem of local history effects.

Selection-Instrumentation Effect A threat due to the combination of selection and instrumentation occurs when changes in a measuring instrument are more likely to be detected in one group than they are in another. Floor or ceiling effects, for instance, could make it difficult to detect changes in behavior from pretest to posttest. If this is more of a problem in one group than in another, a selection-instrumentation effect is present. Shadish et al. (2002) point out that this threat to internal validity is more likely to be a problem the greater the nonequivalence of the groups and the closer the group scores are to the end of the scale. Because Langer and Rodin's groups did not differ on the pretest, and because performance of the groups did not suggest floor or ceiling effects on the measurement scales that were used, this threat to internal validity seems implausible in their study.

Differential Statistical Regression The final threat to internal validity that is not controlled in the nonequivalent control group design is differential statistical regression (Shadish et al., 2002). As we described earlier, regression toward the mean is to be expected when individuals are selected on the basis of extreme scores (e.g., the poorest readers, the workers with the lowest productivity, the patients with the most severe problems). *Differential regression* can occur when regression is more likely in one group than in another. For example, consider a nonequivalent control group design in which the participants with the most serious problems are placed in the treatment group. It is possible, even likely, that regression would occur for this group. The changes from pretest to posttest may be mistakenly interpreted as a treatment effect if regression is more likely in the treatment group than in the control group. Because the groups in the Langer and Rodin study came from the same population and there is no evidence that one group's pretest scores were more extreme than another's, a threat to internal validity due to differential statistical regression is not plausible in their study.

Expectancy Effects, Contamination, and Novelty Effects Langer and Rodin's study could also have been influenced by three additional threats to internal validity that can even affect true experiments—expectancy effects, contamination, and novelty effects. If observers in their study had been aware of the research hypothesis, it is possible that they inadvertently might have rated residents as being better after the responsibility instructions than before. This observer bias, or expectancy effect, appears to have been controlled, however, because all the observers were kept unaware of the research hypothesis. Langer and Rodin were also aware of possible contamination effects. Residents in the control group might have become demoralized if they learned that residents on another floor were given more opportunity to make decisions. In this case, the use of different floors of the nursing home was advantageous; Langer and Rodin (1976) indicate that "there was not a great deal of communication between floors" (p. 193). Thus, contamination effects do not seem to be present, at least on a scale that would destroy the internal validity of the study.

Novelty effects would be present in the Langer and Rodin study if residents on the treatment floor gained enthusiasm and energy as a result of the innovative responsibility-inducing treatment. Thus, this new enthusiasm, rather than treatment residents' increased responsibility, may explain any treatment effects. In addition, the special attention given the treatment group may have produced a Hawthorne effect in which residents on the treated floor felt better about themselves. It is difficult to rule out completely novelty effects or a Hawthorne effect in this study. According to the authors, however, "There was no difference in the amount of attention paid to the two groups" (p. 194). In fact, communications to both groups stressed that the staff cared for them and wanted them "to be happy." Thus, without additional evidence to the contrary, we can conclude that the changes in behavior Langer and Rodin observed were due to the effect of the independent variable, not to the effect of an extraneous variable that the investigators failed to control.

For investigators to decide whether an independent variable "worked" in the context of a particular experiment, they must systematically collect and carefully weigh evidence for and against the interpretation that the treatment caused behavior to change. As Cook and Campbell (1979) explain:

> Estimating the internal validity of a relationship is a deductive process in which the investigator has to systematically think through how each of the internal validity threats may have influenced the data. Then, the investigator has to examine the data to test which relevant threats can be ruled out. In all of this process, the researcher has to be his or her own best critic, trenchantly examining all of the threats he or she can imagine. When all of the threats can plausibly be eliminated, it is possible to make confident conclusions about whether a relationship is probably causal. When all of them cannot, perhaps because the appropriate data are not available or because the data indicate that a particular threat may indeed have operated, then the investigator has to conclude that a demonstrated relationship between two variables may or may not be causal. (pp. 55–56)

The Issue of External Validity

- Similar to internal validity, the external validity of research findings must be critically examined.
- The best evidence for the external validity of research findings is replication with different populations, settings, and times.

We must make the same systematic inquiry into the external validity of a quasi-experiment that we did into its internal validity. What evidence is there that the particular pattern of results is restricted to a particular group of participants, setting, or time? For example, although Langer and Rodin suggest that certain changes be made in the way the elderly are cared for, we might question whether the effectiveness of the responsibility-inducing treatment would hold for all elderly residents, for all types of nursing facilities, and at different times. That the particular nursing home selected by Langer and Rodin (1976) was described as "rated by the state of Connecticut as being among the finest care units" (p. 193) suggests that the residents, facilities, and staff might be different from those found in other facilities. For instance, if residents at this particular nursing home were relatively more independent before coming to the home than residents at other homes (perhaps because of differences in socioeconomic status), then the changes experienced upon moving into a home might have had greater impact on them. Consequently, the opportunity to be more independent of staff might be more important to these residents relative to residents in other homes. Similarly, if staff members at this home were more competent than those at other homes, they might be more effective in communicating with the residents than would the staff members at other homes.

In the last analysis, the investigator must be ready to *replicate* an experimental finding with different populations, settings, and times in order to establish external validity. The deductive process applied to questions of internal validity must also be used to examine a study's external validity. Moreover, *we must be ready to live with the fact that one study is not likely to answer all questions about a research hypothesis*.

Interrupted Time-Series Designs

- In a simple interrupted time-series design, researchers examine a series of observations both before and after a treatment.
- Evidence for treatment effects occurs when there are abrupt changes (discontinuities) in the time-series data at the time treatment was implemented.
- The major threats to internal validity in the simple interrupted time-series design are history effects and changes in measurement (instrumentation) that occur at the same time as the treatment.

Key Concept

A second quasi-experiment, a **simple interrupted time-series design,** is possible when researchers can observe changes in a dependent variable for some time before and after a treatment is introduced (Shadish et al., 2002). The essence of this design is the availability of periodic measures before and after a treatment has been introduced. The simple interrupted time-series design can be outlined in the following way:

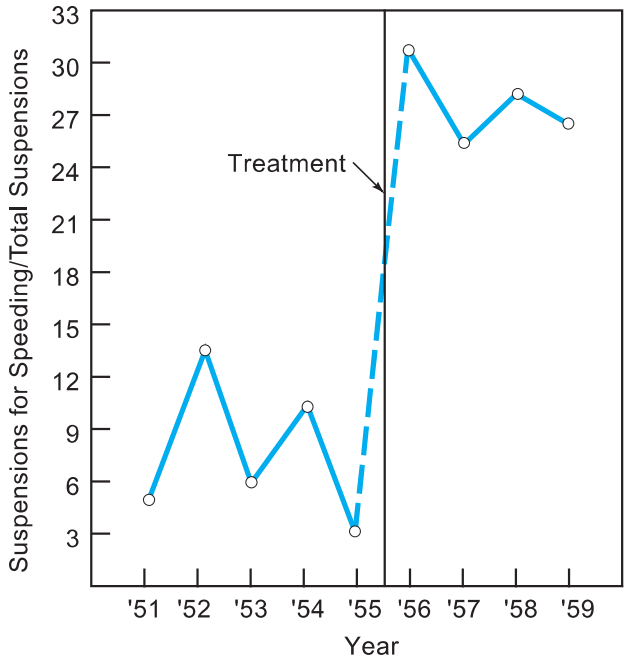
$$O_1\ O_2\ O_3\ O_4\ O_5\ X\ O_6\ O_7\ O_8\ O_9\ O_{10}$$

The simple interrupted time-series design can be used to assess the effect of a treatment in situations such as when a new product has been introduced, a new social reform instituted, or a special advertising campaign begun. Campbell (1969) investigated the effect of a social policy change in Connecticut in the mid-1950s. The governor had ordered a crackdown on speeding, and Campbell made use of an interrupted time-series design to determine the effect of this order on traffic fatalities. Campbell was able to obtain a wealth of archival data to use as pretreatment and posttreatment measures because statistics related to traffic accidents are regularly kept by state agencies. Besides number of fatalities, Campbell looked at number of speeding violations, number of drivers having their licenses suspended, and other measures related to driving behavior. Figure 10.5 shows the percentage of suspensions of licenses for speeding (as a percentage of all license suspensions) before and after the crackdown. There is a clear discontinuity in the graph that coincides with the onset of the treatment. This discontinuity provides evidence for an effect of the treatment. Indeed, *a discontinuity in the time series is the major evidence of an effect of treatment.*

As Campbell points out, only abrupt changes in the time-series graph can be interpreted because gradual changes are indistinguishable from normal fluctuations over time. Unfortunately, changes often are not nearly so dramatic as those seen in Figure 10.5. In fact, Campbell's analysis of traffic fatalities over the same time period did reveal evidence for an effect of the crackdown, but the change in traffic fatalities was not as abrupt as that associated with suspension of drivers' licenses (see Campbell, 1969, Figure 2).

A variation of the interrupted time-series design was used to assess the effect of avoiding the "dread risk of flying" following the terrorist attacks on the United States on September 11, 2001 (Gigerenzer, 2004). The rationale for this study was as follows. People tend to fear "dread risks," which are defined

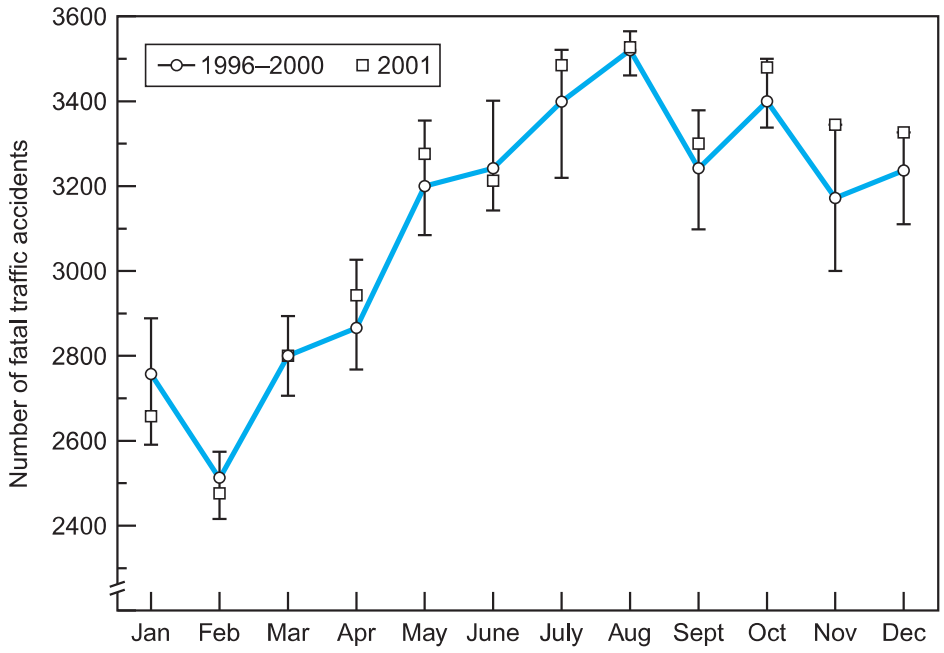
FIGURE 10.5 Suspensions of licenses for speeding, as a percentage of all suspensions. (From Campbell, 1969.)



as "low-probability, high-consequence events, such as the terrorist attacks on September 11, 2001" (Gigerenzer, 2004, p. 286). If Americans, in order to avoid the dread risk of flying, instead drove to their destinations, then an increase in traffic fatalities would be expected. To test this hypothesis, Gigerenzer (2004) examined data from the U.S. Department of Transportation for the 3 months following September 11, 2001. Data also were analyzed for the 5 years prior to that date. The mean number of fatalities in these preceding years was compared with the numbers after September 11, 2001. The results of this analysis are seen in Figure 10.6.

The graph shows fatal traffic accidents for all 12 months of the year for both the preceding 5 years (circles depict the means in the graphed line) and for the year 2001 (depicted by squares). In addition, the highest and lowest values for each month in the preceding 5 years are shown (the bars around each mean). The data for fatalities during October, November, and December reveal that in the year 2001, the number of fatal traffic accidents was as high as or higher than the highest value for the preceding 5 years. On the basis of these data (and statistical analyses), Gigerenzer (2004) was able to conclude that traffic fatalities increased by 353 people in October, November, and December 2001. He attributed this increase to Americans' dread fear of flying following the events of 9/11. Gigerenzer compared this increase of 353 deaths to the 266 passengers and crew who were killed in the four plane crashes (and of course many more on the ground). The researcher suggested that "if the public were better informed about psychological reactions to catastrophic events, and the potential risk of avoiding risk," perhaps this "psychologically motivated toll" could be prevented (p. 287).

FIGURE 10.6 Number of fatal traffic accidents in the United States in 1996 through 2000, versus 2001. The graphed line represents the means for the years 1996–2000; the bars around the means indicate the lowest and highest values for those years. The squares indicate the numbers of fatal traffic accidents for each month in 2001. (From Gigerenzer, 2004.)



Although results of the interrupted time-series design and other quasi-experimental designs are sometimes able to be interpreted on the basis of visual inspection (see, for example, Figure 10.5), statistical analyses are often needed. Gigerenzer (2004), for example, used a chi-square test of statistical significance to demonstrate that there was a statistically significant increase in the frequency of traffic fatalities following September 11, 2001, compared to the preceding 5 years. We previously mentioned the chi-square test in Chapter 5. In other situations, more sophisticated analyses may need to be conducted (e.g., Michielutte, Shelton, Paskett, Tatum, & Velez, 2000). For more information, refer to Shadish et al.'s (2002) text, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*.

Campbell and Stanley (1966) summarize the problem facing researchers using the simple interrupted time-series design: "The problem of internal validity boils down to the question of plausible competing hypotheses that offer likely alternate explanations of the shift in the time series other than the effect of X" (p. 39). An effect of history is the main threat to internal validity in this type of design (Shadish et al., 2002). For instance, is it possible that some factor other than avoiding "dread risk" was responsible for the increase in fatal traffic accidents in the last months of 2001 (see Figure 10.6)?

Particularly threatening to the internal validity of the time-series design are influences of a cyclical nature, including seasonal variation (Cook & Campbell,

1979). For instance, when analyzing the effect of the Connecticut governor's crackdown on speeding, Campbell (1969) gathered data from neighboring states to rule out possible regional trends due to weather in order to strengthen his case for the effect of this particular social policy change.

Instrumentation must also be considered a threat to internal validity in the simple interrupted time-series design (Shadish et al., 2002). When new programs or new social policies are instituted, for example, there are often accompanying changes in the way records are kept or in the procedures used to collect information. A program intended to reduce crime may lead authorities to modify their definitions of particular crimes or to become more careful when observing and reporting criminal activities. Nevertheless, for an instrumentation threat to be plausible, the changes in instrumentation must be shown to have occurred at *exactly* the time as the intervention (Campbell & Stanley, 1966). Threats to internal validity due to maturation, testing, and regression are controlled in the simple interrupted time-series design. None of these threats can be ruled out when only a single pretest and posttest measure is available. These threats are nearly eliminated, however, by the presence of multiple observations both before and after treatment. For example, an effect of maturation would not normally be expected to show a sharp discontinuity in the time series, although this might be possible in some situations (Campbell & Stanley, 1966).

Threats to external validity in the simple interrupted time-series design must be examined carefully. When pretreatment observations of behavior are based on multiple tests, then it is very likely that an effect of the treatment may be restricted to those individuals who have had these multiple test experiences. Moreover, the interrupted time-series design generally involves testing only a single group that has not been randomly selected. This aspect of the design leaves open the possibility that the results are limited to people with characteristics similar to those who took part in the study.

Time Series with Nonequivalent Control Group

- In a time series with nonequivalent control group design, researchers make a series of observations before and after treatment for both a treatment group and a comparable comparison group.

Key Concept

The internal validity of the interrupted time-series design can be enhanced greatly by including a control group following the procedures we described earlier for the nonequivalent control group design. For the **time series with nonequivalent control group design** the researcher must find a group that is comparable to the treatment group and that allows a similar opportunity for multiple observations before and after the time that the treatment is administered to the experimental group. This design is outlined as follows:

$$O_1\ O_2\ O_3\ O_4\ O_5\ \ X\ \ O_6\ O_7\ O_8\ O_9\ O_{10}$$

$$O_1\ O_2\ O_3\ O_4\ O_5 \qquad O_6\ O_7\ O_8\ O_9\ O_{10}$$

As before, a dashed line is used to indicate that the control group and the experimental group were not randomly assigned. The interrupted time series with nonequivalent control group design permits researchers to rule out many threats due to history. As was mentioned earlier, Campbell (1969) used traffic-fatality data obtained from neighboring states to provide a comparison with traffic-fatality data following the crackdown on speeding in Connecticut. Although traffic fatalities in Connecticut showed a decline immediately following the crackdown, data from comparable states did not exhibit any such decline. This finding tends to rule out claims that the decrease in traffic fatalities in Connecticut were due to factors such as favorable weather conditions, improved automobile design, or any other factors that were likely shared by Connecticut and the neighboring states.

Program Evaluation

- Program evaluation is used to assess the effectiveness of human service organizations and provide feedback to administrators about their services.
- Program evaluators assess needs, process, outcome, and efficiency of social services.
- The relationship between basic research and applied research is reciprocal.
- Despite society's reluctance to use experiments, true experiments and quasi-experiments can provide excellent approaches for evaluating social reforms.

Organizations that produce goods have a ready-made index of success. If a company is set up to make microprocessors, its success is ultimately determined by its profits from the sale of microprocessors. At least theoretically, the efficiency and effectiveness of the organization can be easily assessed by examining the company's financial ledgers. Increasingly, however, organizations of a different sort play a critical role in our society. Because these organizations typically provide services rather than goods, Posavac (2011) refers to them as human service organizations. For example, hospitals, schools, police departments, and government agencies provide a variety of services ranging from emergency room care to fire prevention inspections. Because profit-making is not their goal, some other method must be found to distinguish between effective and ineffective agencies. One useful approach to assessing the effectiveness of human service organizations is program evaluation.

Key Concept

According to Posavac (2011), **program evaluation** is

> a methodology to learn the depth and extent of need for a human service and whether the service is likely to be used, whether the service is sufficiently intensive to meet the unmet needs identified, and the degree to which the service is offered as planned and actually does help people in need at a reasonable cost without unacceptable side effects. (p. 1)

The definition of program evaluation includes several components; we will take up each of these components in turn. Posavac emphasizes, however, that the overarching goal of program evaluation is *to provide feedback regarding human service activities*. Program evaluations are designed to provide feedback to the administrators of human service organizations to help them decide what

services to provide to whom and how to provide them most effectively and efficiently. Program evaluation is an integrative discipline that draws on political science, sociology, economics, education, and psychology. We are discussing program evaluation at the end of this chapter on research in natural settings because it represents perhaps the largest-scale application of the principles and methods we have been describing throughout this book.

Posavac (2011) identifies four questions that are asked by program evaluators. These questions are about needs, process, outcome, and efficiency. An assessment of *needs* seeks to determine the unmet needs of the people for whom an agency might provide a service. Consider, for example, a city government that has received a proposal to institute a program of recreational activities for senior citizens in the community. The city would first want to determine whether senior citizens actually need or want such a program. If the senior citizens do want such a program, the city would further want to know what kind of program would be most attractive to them. The methods of survey research are used extensively in studies designed to assess needs. Administrators can use the information obtained from an assessment of needs to help them plan what programs to offer.

Once a program has been set up, program evaluators may ask questions about the *process* that has been established. Observational methods are often useful in assessing the processes of a program. Programs are not always implemented the way they were planned, and it is essential to know what actually is being done when a program is implemented. If the planned activities were not being used by the senior citizens in a recreational program designed specifically for them, it might suggest that the program was inadequately implemented. An evaluation that provides answers to questions about process, that is, about how a program is actually being carried out, permits administrators to make adjustments in the delivery of services in order to strengthen the existing program (Posavac, 2011).

The next set of questions a program evaluator is likely to ask involves the *outcome*. Has the program been effective in meeting its stated goals? For example, do senior citizens now have access to more recreational activities, and are they pleased with these activities? Do they prefer these particular activities over other activities? The outcome of a neighborhood-watch program designed to curb neighborhood crime might be evaluated by assessing whether there were actual decreases in burglaries and assaults following the implementation of the program. It is possible to use archival data like those described in Chapter 4 to carry out evaluations of outcome. For example, examining police records in order to document the frequency of various crimes is one way to assess the effectiveness of a neighborhood-watch program. Evaluations of outcome may also involve both experimental and quasi-experimental methods for research in natural settings. An evaluator may, for example, use a nonequivalent control group design to assess the effectiveness of a school reform program by comparing students' performance in two different school districts, one with the reform program and one without.

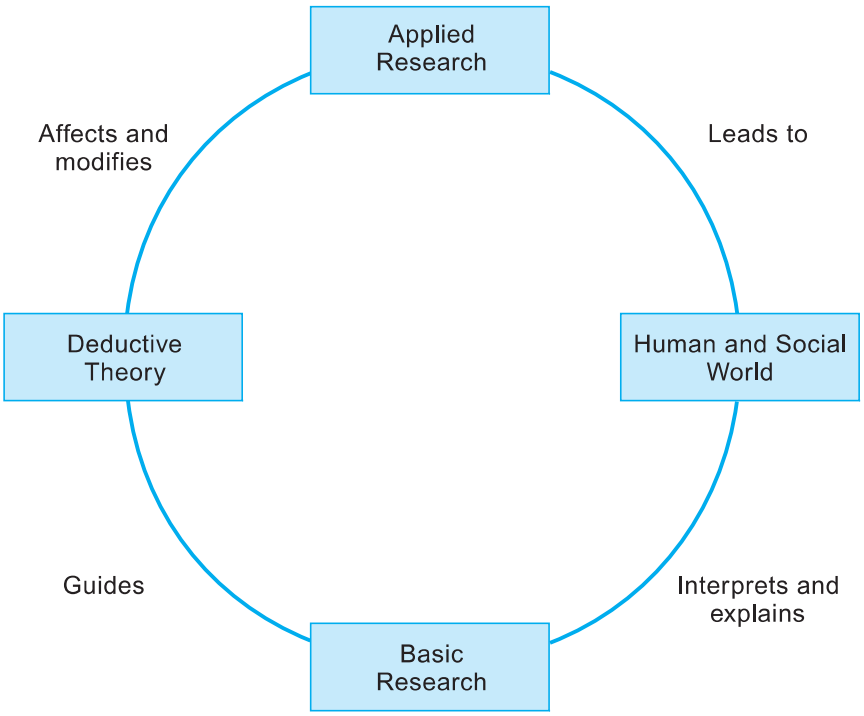
The final questions evaluators might ask are about the *efficiency* of the program. Most often, questions about efficiency relate to the cost of the program. Choices often have to be made among possible services that a government or other institution is capable of delivering. Information about how successful a program is (outcome evaluation) and information about the program's cost

(efficiency evaluation) are necessary if we want to make informed decisions about continuing the program, how to improve it, whether to try an alternative program, or whether to cut back on the program's services.

Earlier in this chapter and in Chapter 2 we described differences between *basic and applied research.* Program evaluation is perhaps the extreme case of applied research. The purpose of program evaluation is practical, not theoretical. Nevertheless, even in the context of blatantly practical goals, a case can be made for a reciprocal relationship between basic and applied research (see Box 10.1). One such model of this relationship is illustrated in Figure 10.7. The idea is that each domain of research serves the other in an ongoing circular way. Specifically, basic research provides us with certain abstractions (e.g., scientifically based principles) that express certain regularities in nature. When these principles are examined in the complex and "dirty" world where they supposedly apply, new complexities are recognized and new hypotheses are called for. These new complexities are then tested and evaluated in the lab before being tried out again in the real world.

The work of Ellen Langer serves as a concrete example of this circular relationship (see Salomon, 1987). She identified a decline in elderly people's health once they entered nursing homes (see Langer, 1989; Langer & Rodin, 1976, described in this chapter). These naturalistic observations led her to develop a theory of mindfulness, which she has tested under controlled experimental conditions and which has implications for more general theories of cognitive development and of education (see, for example, Langer, 1989, 1997; Langer & Piper, 1987). The theory provides a guide for her applied work—designing new models of nursing homes. Tests of the practical effects of changes in the care

FIGURE 10.7 Model illustrating reciprocal relationship between basic and applied research. (From Salomon, 1987, p. 444.)



given by nursing homes on the residents' health and well-being will undoubtedly lead to modifications of her theory of mindfulness.

According to Campbell (1969), it is important for public officials involved with social experiments to emphasize the importance of the problem rather than the importance of the solution. Instead of pushing for one certain cure-all (which, in most cases, has little opportunity for success), officials must be ready to execute reform in a manner that permits the clearest evaluation and must be prepared to try different solutions if the first one fails. Public officials must, in other words, be ready to use the experimental method to identify society's problems and to determine effective solutions.

Campbell's (1969) idea that social reforms and experimental methods be routinely brought together has had some impact on social policymakers, but it is still underutilized (see Berk et al., 1987, and Box 10.3). The reasons are some of the same ones initially identified by Campbell. Nevertheless, without

BOX 10.3

REMOVING THE "CLOAK OF IGNORANCE" FROM SOCIETY'S EXPERIMENTS

Perhaps the greatest difference between basic research and program evaluation lies in the political and social realities surrounding program evaluation. Governments at both local and national levels regularly propose, plan, and execute various types of social reforms. Tax relief programs, work incentive programs, educational reforms, police reforms, and medical care for senior citizens are just a few of the types of social reform programs that a government might initiate. Unfortunately, as the late Donald Campbell (1969), a past president of the American Psychological Association, pointed out, the outcome of these social reforms often cannot be meaningfully evaluated. Did a change in police techniques lead to less crime? Are more elderly people gaining access to public transportation after a reduction in fares? Does a work incentive program take more people off the unemployment rolls? The answers to such questions often cannot be found, said Campbell, because most social reforms are instituted in a political climate that is not ready for hardheaded evaluation. What public official, for instance, wants to be associated with a program that failed? As Campbell suggested, there is "safety under the cloak of ignorance" (pp. 409–410). Furthermore, many social reforms are begun under the assumption that they are certain to be successful. Otherwise, why spend all that tax money? For many public administrators it is advantageous to leave that assumption in people's minds rather than face the truth about what happened.

Campbell put forth the argument that

> the United States and other modern nations should be ready for an experimental approach to social reform, an approach in which we try out new programs, in which we learn whether or not these programs are effective, and in which we retain, imitate, modify, or discard them on the basis of apparent effectiveness on the multiple imperfect criteria available. (p. 409)

Social scientists need to convince administrators to use true experiments, if at all possible, or quasi-experiments at the very least, when instituting new social programs. For example, a randomization procedure, perhaps based on public lottery, could be used to decide which group receives a pilot program or gains access to scarce resources. Groups not receiving the program or the available resources would become comparison groups. The effect of a social "treatment" could then be meaningfully evaluated. At present, decisions regarding who gets what are often influenced by particular interest groups—as the result of intense lobbying, for example—or made on the basis of political favoritism.

social experimentation, especially that which makes use whenever possible of randomized field experiments, policymakers and the community at large may believe a treatment works when it doesn't or vice versa. Such incorrect decisions lead us to allocate money and resources to ineffective programs.

Not too many years ago, a show called "Scared Straight" was aired on national television. It described a juvenile education program implemented at Rahway State Prison in New Jersey. The program involved taking youthful offenders into a prison to meet with selected convicts from the inmate population. The goal was to inform juveniles about the reality of prison life and, thereby, the program leaders hoped, dissuade them from further illegal activity. Unsubstantiated claims were made for the effectiveness of the program, including some suggesting a success rate as high as 80% to 90% (see Locke, Johnson, Kirigin-Ramp, Atwater, & Gerrard, 1986). The Rahway program is just one of several similar programs around the country. But do these programs really work?

Several evaluation studies of the exposure-to-prison programs produced mixed results, including positive findings, findings of no difference between control and experimental participants, as well as results suggesting that the program may actually increase juvenile crime among some types of delinquents. There is a possibility that less hardened juvenile offenders may increase their criminal activity after meeting the prisoners. It has been suggested that, because these less hardened offenders have recently begun a lifestyle wherein they are being recognized and reinforced by their peers for their toughness, this image is also reinforced by the tough image often projected by the prisoners. On the other hand, hardened juvenile offenders, who have achieved a level of status among their peers for some period of time, may be more threatened by the prospects of prison life because it would mean loss of that status (see Locke et al., 1986).

Attempts to evaluate the effectiveness of this significant social program provide good examples of the difficulties inherent in evaluation research: the difficulty of randomly assigning participants, of getting administrators to cooperate with experimental procedures, and of dealing with loss of participants during the evaluation. Nevertheless, program evaluation based on sound experimental methodology offers policymakers at all levels (institution, community, city, state, federal) the information that can help them make informed choices among possible treatments for social problems. Because resources inevitably are in short supply, it is critical that resources be put to the best possible use. *Our hope is that your knowledge of research methods will allow you to participate knowledgeably and perhaps contribute constructively to the ongoing debate concerning the role of experimentation in society.*

Summary

Experimentation in natural settings differs in many ways from experimentation in psychology laboratories. The reasons for doing experiments in natural settings include testing the external validity of laboratory findings and assessing the effects of "treatments" aimed at improving conditions under which people work and live.

Many social scientists have argued that society must be willing to take an experimental approach to social reform—one that will allow the clearest evaluation of the effectiveness of new programs. In many situations (for instance, when available resources are scarce), true experiments involving randomization of individuals to treatment and no-treatment conditions are recommended. However, if a true experiment is not feasible, quasi-experimental procedures are the next best approach. Quasi-experiments differ from true experiments in that fewer plausible rival hypotheses for an experimental outcome are controlled. When specific threats to the internal validity of an experiment are not controlled, then the experimenter, by logically examining the situation and by collecting additional evidence, must seek to rule out these threats to internal validity.

A particularly strong quasi-experimental procedure is the nonequivalent control group design. This procedure generally controls for all major threats to internal validity except those associated with additive effects of (1) selection and history, (2) selection and maturation, (3) selection and instrumentation, and (4) threats due to differential statistical regression. In addition to the major threats to internal validity, an experimenter must be sensitive to possible contamination resulting from communication between groups of participants. Problems of experimenter expectancy effects (observer bias); questions of external validity; and novelty effects, including the Hawthorne effect, are potential problems in all experiments, whether conducted in the laboratory or in the field.

When it is possible to observe changes in a dependent measure before and after a treatment is administered, one can carry out a simple interrupted time-series design. The researcher using this design looks for an abrupt change (discontinuity) in the time series that coincides with the introduction of the treatment. The major threat to internal validity in this design is history—some event other than the treatment may have been responsible for the change in the time series. Instrumentation also can be a problem, especially when the treatment represents a type of social reform that may lead to changes in the way records are kept or data collected. By including a control group that is as similar as possible to the experimental group, one can strengthen the internal validity of a simple time-series design. A time series with nonequivalent control group, for example, controls for many possible history threats.

A particularly important goal of research in natural settings is program evaluation. Professionals other than psychologists (such as educators, political scientists, and sociologists) are often involved in this process. Types of program evaluation include assessment of needs, process, outcome, and efficiency. Perhaps the most serious constraints on program evaluation are the political and social realities that surround it. The reluctance of public officials to seek an evaluation of social reforms is often an obstacle to be overcome. Nevertheless, social scientists have called on program evaluators to make themselves available to human services organizations. By answering this call, we may help change society in a way that will bring the most effective services to those most in need.

Key Concepts

threats to internal validity 315
history 315
maturation 315
testing 316
instrumentation 316
regression 316
subject attrition 317
selection 317
contamination 319
novelty effects 319
quasi-experiments 321
nonequivalent control group design 323
simple interrupted time-series design 332
time series with nonequivalent control group design 335
program evaluation 336

Review Questions

1 Identify two reasons why it might be especially important to carry out experiments in natural settings.
2 Explain how laboratory experiments and those in natural settings differ in control, external validity, goals, and consequences.
3 Describe the three distinguishing characteristics of true experiments, and identify how the independent variable can be defined in terms of these characteristics.
4 What obstacles do researchers have to overcome when they try to carry out experiments in natural settings?
5 Identify two procedures that permit researchers to assign participants randomly to conditions while still giving all participants access to the experimental treatment.
6 Describe and explain the consequences of the three ways in which participants in a control group might respond when contamination occurs.
7 Explain how novelty effects, including the Hawthorne effect, may influence a researcher's interpretation of the effectiveness of an experimental treatment.
8 What do Cook and Campbell (1979) consider the best test of external validity?
9 Explain why it is essential to use a pretest in the nonequivalent control group design.
10 Explain how one threat to internal validity is controlled in the nonequivalent control group design, and describe a threat to internal validity that is not controlled in this design.
11 Identify two reasons why we cannot conclude that the treatment and control groups in a nonequivalent control group design are equivalent even when the pretest scores are the same for both groups.
12 Explain the difference between a history threat to internal validity and what is called a "local history effect" in the nonequivalent control group design.
13 What is the major evidence for an effect of the treatment in a simple interrupted time-series design, and what are the major threats to internal validity in this design?
14 Explain how the addition of a nonequivalent control group to a simple interrupted time-series design reduces the threat to the internal validity of the design.
15 Describe what type of information is being sought when evaluators ask each of the four questions typically addressed in program evaluation.

CHALLENGE QUESTIONS

1 A quasi-experiment was used to determine whether multimedia instruction is effective. Two sections of introductory psychology were taught by the same instructor, both in the early afternoon. In one section (the treatment group), the instructor used multimedia instruction. In the other section, the instructor covered the same material but did not use multimedia instruction. Students did not know when they registered for the course whether multimedia instruction would be used, but the students were not randomly assigned to sections. Students' knowledge of the course material was assessed using two forms of a comprehensive introductory psychology test. The comprehensive test can be considered a reliable and valid test that can be used to compare the effectiveness of the instruction in the two sections. The students in both sections were tested on the second day of class (the pretest) and at the final (the posttest). Different forms of the test were used at the pretest and at the posttest.

- **A** What quasi-experimental design is used in the study?
- **B** The instructor initially considered doing a true experiment rather than a quasi-experiment. Comment critically on the fairness of random assignment if you were arguing in favor of doing a true experiment to test the effectiveness of multimedia instruction.
- **C** Explain why the quasi-experimental design used by the instructor is more effective than if the instructor had tested only students who had received multimedia instruction. Identify one threat to internal validity that was controlled in this study that would not have been controlled if only students who received multimedia instruction had been tested.

2 A psychologist published a book describing the effects of divorce on men, women, and children. She was interested in the effects of divorce that occurred 10 years after the divorce. She found that even 10 years after a divorce half of the women and one third of the men were still intensely angry. Although half the men and women described themselves as happy, 25% of the women and 20% of the men remained unable to "get their lives back on track." In only 10% of the divorced families did both the former husbands and wives have happy, satisfying lives a decade later. Finally, more than half of the children of divorce entered adulthood as underachieving and self-deprecating men and women. These findings were based on a 15-year study of 60 divorced couples and their 131 children living in Marin County, California (an affluent suburban area including mostly well-educated people). Explain how the use of a quasi-experimental design would have been helpful in order to specify which of the reported results are due to the effects of divorce.

3 The police force of a large city had to decide between two different approaches to keeping the officers on the force informed about the changes in laws. An enlightened administrator of this force decided to put the two approaches to test in a research study. She decided to do a true experiment and assigned 30 officers randomly to each of the two programs for a period of 6 months. At the end of this time, all the officers who successfully completed the training under the two approaches were given a final test on their knowledge of the law. The 20 officers who completed Program A showed a reliably higher mean score on this test than did the 28 officers who completed Program B. The administrator wisely chose not to accept these results as decisive evidence of the effectiveness of the two programs. Using only the data reported in this problem, explain why she made this decision. Next, explain how her decision would have been different if only 20 officers completed both programs (from the original 30 assigned to each) and there was still a sizable difference favoring Program A. Be sure to mention any limitations on the conclusions she could reach concerning the overall effectiveness of these programs.

4 A small undergraduate college with a new physical-fitness center decided to introduce a health enhancement program for faculty and staff. The program is designed to take one semester to complete with three 1-hour sessions per week. Comment critically on each of the following questions regarding the evaluation of this program.

- **A** How might an assessment of needs have played a role in planning the program?
- **B** What questions about the process of the program would be useful once the program was under way to help ensure that the evaluation of the outcome of the program could be interpreted appropriately?
- **C** Explain how you would test the effectiveness of the proposed program if it were not possible to do a true experiment.

Answer to Stretching Exercise

1 History is a threat when an event other than the treatment can explain the participants' improvement. For example, participants may have read self-help books, tried herbal supplements, talked to friends or pastors, or experienced any number of potentially beneficial "treatments." Any of these other events may have caused the depression to improve, rather than the psychologist's treatment.

2 Maturation occurs when participants naturally change over time. One of the things we know about depression is that it tends to improve with time. Therefore, the participants' improvement may reflect natural decreases in depression over time, rather than the effect of the treatment.

3 A testing threat occurs when a first administration of a test influences subsequent testing. In this study, participants may have remembered their earlier responses on the depression measure and, perhaps in an effort to demonstrate they improved, chose responses that indicated less depression at posttest (even if they didn't feel less depressed). An instrumentation threat occurs when the measure used to assess thoughts, feelings, and behavior changes over time. Because the same questionnaire was used for both the pretest and posttest, this threat is less likely.

4 Statistical regression is possible when participants are selected because they are extreme on a pretest measure. In this study, participants were selected because they were depressed—they scored high on a measure of depression. It's possible that the lower scores at posttest indicated improvement because of statistical regression to the mean, not because of the effects of treatment.

Answer to Challenge Question 1

A The nonequivalent control group design was used in this study.

B Students may perceive random assignment to the two sections as unfair because they would not have a choice about which section they would take. If we do not know whether multimedia instruction is effective, then random assignment is the best and fairest method to determine whether multimedia instruction is effective.

C If only the students who had received multimedia instruction had been tested, the design of the study would have been a single group pretest-posttest design. There are several threats to the internal validity of a pretest-posttest single group design. That is why it is referred to as a pre-experimental design or a bad experiment. One possible threat in this study is due to testing; that is, students often improve from an initial test in a course to the second test because they gain familiarity with the testing procedure and the instructor's expectations. This improvement would be expected to occur even if multimedia instruction had not been used. The nonequivalent control group design in this study controls for this threat because any increase in test scores due to testing effects would likely be the same for both groups. A *greater* increase from the pretest to the posttest for the group given multimedia instruction, relative to the control group, can be interpreted as an effect of the instruction.