Quasi-Experimental Designs and Applied Research

11

PREVIEW & CHAPTER OBJECTIVES

In this chapter, we consider a type of research design that, like an experiment, includes independent and dependent variables but involves a situation in which research participants cannot be randomly assigned to groups. Because the absence of random assignment means causal conclusions cannot be made, whereas they can be made with some degree of confidence in a purely experimental study, this design is called quasi-experimental ("almost" experimental). One type of design you have already encountered (in Chapter 5) is an example of a quasi-experimental design—any study having subject variables. The inability to randomly assign often (but not always) occurs in applied research that takes place outside of the lab, so one focus of the chapter will be applied research, a strategy you first encountered in Chapter 3, when it was contrasted with basic research. You will learn that applied research represents a strong tradition in American experimental psychology and reflects the core American value of pragmatism. Program evaluation is a form of applied research that uses a variety of strategies to examine the effectiveness of programs designed to help people. Program evaluation is especially likely to use qualitative analysis. When you finish this chapter, you should be able to:

- Identify the dual functions of applied research.
- Understand why applied psychology has always been an important element in American psychology.
- Define translational research and explain how psychological research can translate into applied settings.
- Identify the design and ethical problems associated with applied research, especially if that research occurs outside of the laboratory.
- Identify the defining feature of a quasi-experimental design, and recognize which designs appearing in earlier chapters were quasi-experimental.
- Describe the features of a nonequivalent control group design, and understand why this design is necessarily confounded.
- Understand why matching nonequivalent groups on pretest scores can introduce a regression artifact.
- Describe the features of interrupted time series designs, and understand how they can be used to evaluate trends.
- Describe several variations on the basic time series design.
- Describe the strategies for completing a needs analysis in a program evaluation project.

- Understand the purposes and the procedures involved in formative evaluation, summative evaluation, and cost-effectiveness evaluation.
- Identify and describe the ethical problems that often accompany program evaluation research.

As mentioned at the beginning of this text, we would like nothing more than to see you emerge from this methods course with a desire to contribute to our knowledge of behavior by becoming a research psychologist. Our experiences as teachers in this course tell us that some of you indeed will become involved in research, but most of you won't. Many of you will become professional psychologists of some kind, however, working in fields that focus on the development, implementation, and assessment of programs to improve the human condition. For example, as a health psychologist, you might find yourself involved in program to improve the physical and psychological well-being of clients; as a school psychologist, you might be asked to evaluate programs designed to improve student learning; or as an industrial-organizational psychologist, you might help develop programs to improve worker productivity and job satisfaction. As such, you will encounter the worlds of applied research and program evaluation. You may discover you need to do things like:

- Read, comprehend, and critically evaluate research literature on the effectiveness of a program your agency is thinking about implementing.
- Help plan a new program by informing (tactfully) those who are less familiar with research design about the adequacy of the evaluation portion of their proposal.
- Participate in an agency self-study in preparation for an accreditation process.

And if your agency's director finds out you took this course, you might even be asked to design and lead a study to evaluate an agency program.

Beyond the Laboratory

You first learned to distinguish between basic and applied research in the opening pages of Chapter 3. To review, the essential goal of basic research in psychology is to increase our core knowledge about human behavior and mental processes. The knowledge might eventually have a practical application but that outcome is not the prime motivator; knowledge is valued as an end in itself. In contrast, applied research is designed primarily to increase our knowledge about a particular real-world problem, with an eye toward directly solving it. A second distinction between basic and applied research is that while basic research usually takes place in a laboratory, applied research is often conducted in clinics, social service agencies, jails, government agencies, and business settings. There are many exceptions, of course. Some basic research occurs in the field, and some applied research takes place in a lab.

To give you a sense of the variety of applied research, consider these 2015 titles from the prominent *Journal of Experimental Psychology: Applied:*

- Humanizing machines: Anthropomorphization of slot machines increases gambling (Riva, Sacchi, & Brambilla, 2015).
- Goal-oriented training affects decision-making processes in virtual and simulated fire and rescue environments (Cohen-Hatton & Honey, 2015).
- The interactive effects of affect and shopping goal on information search and product evaluations (Fangyuan, Wyer, & Shen, 2015).

These titles illustrate two features of applied research. First, following from our definition, the studies clearly focus on easily recognizable problems (gambling, decision-making in emergency situations, and shopping behavior). Second, the titles demonstrate that, while the prime goal of applied research is problem solving (e.g., how to get firefighters to make good decisions in emergency situations), these studies also further our knowledge of basic psychological processes (e.g., decision making).

Indeed, there is a close connection between basic and applied research, as illustrated by growing field of *translational research*. In Chapter 1, we defined translational research as research that is done for both better understanding of a particular phenomenon as well as for its application to promote physical and psychological well-being. While basic research may serve as the "engine of discovery," driving innovation and deeper understanding of human functioning, it is also important that basic research results apply to situations that enable users of research to inform their practice. Further, to best inform therapeutic interventions, basic research findings need to be *translated* and tested in clinical situations. The National Institutes of Health (NIH) has recognized this need and has made translational research a priority (Woolf, 2008). Broadly speaking, translational research has been called "bench-to-bedside" approaches for translating basic research into interventions and treatments for individuals. In psychology, it has been considered a type of research that can help bridge the science-practice gap (Tashiro & Mortensen, 2006).

Virtually, all applied research has the dual function of addressing applied problems directly and providing evidence of basic psychological phenomena that influence theory development. Furthermore, applied research often is rooted in theories and research findings derived from basic research. One illustration of these points comes from the following example of applied research on the impact of nutritional labeling on perceptions of food health and food choice.

Research Example 33—Applied Research

In Chapter 3, we introduced you to distinction between basic and applied research. On the one hand, basic research can provide us with more knowledge and understanding about various psychological constructs, like attention and memory. On the other hand, applied research can make use of basic research findings and develop empirical studies to both understand and attempt to solve real-world problems, like attention to food labels and making healthy food choices. In a study by Trudel, Murray, Kim, and Chen (2015), basic research is used as a basis for conducting applied research on food preferences and food choices based on consumers' attention to the color-coding of nutrition labels.

In a series of four experiments, Trudel et al. (2015) relied on past basic research demonstrating that traffic light color-coded (TLC) nutrition labels can be useful decision aids for consumers. Green labels should signal "go" for consumers to consume the food, and yellow and red should signal increasing caution in consumption. They wanted to see if such TLC labels are related to how consumers evaluate how healthy a food product is and whether consumers would choose to eat those foods. They also used prior theoretical work on self-regulation of eating behaviors to guide their hypotheses about whether those who were watching their diet versus those that were not would respond differently to TLC labels. Self-regulation is the process by which we try to control our thoughts, emotions, and impulses. With regard to eating behavior, previous research has shown that people rely on external cues to self-regulate their impulses and food intake (Trudel & Murray, 2011). Trudel et al. predicted that dieters would affected by external cues (i.e., TLC food labels) differently than non-dieters, which should result in different processing of nutrition information, different food preferences, and different food choices.[1]

[1] Participants answered various questions on a survey, one of which was "Are you currently watching your weight?" (Trudel et al., 2015, p. 258). If participants answered 'yes' to this item, they were classified as dieters, and those who answered 'no' were classified as non-dieters.

Dieters and non-dieters were shown a photo and verbal description of various food products, such as the "Chicken Marbella Sandwich" (Trudel et al., 2015). In the experimental condition, participants saw a nutrition label with various rows color-coded in red, yellow, and green. In the control condition, participants saw a regular black-and-white nutrition label. Then, participants rated how healthy the food items was on a 9-point scale. In Experiment 1, the color coding of the nutrition labels included red (3 rows), yellow (1 row), and green (3 rows). Nondieters rated items as healthy regardless of whether a TLC label was used, whereas dieters rated food items with TLC labels as less healthy than items without TLC labels. Experiment 2 *replicated* the results of Experiment 1 using labels that were color coded either predominantly red or predominantly green. The authors suggested that dieters used the TLC labeling in a way that allowed them to more deeply process nutrition information provided in the labels, leading to more nuanced judgments about the healthiness of the foods. To test this, they conducted a third experiment in which they tested participants' memories for nutrition label information. They found that dieters recalled significantly more information from the food labels when TLC color coding was used than when it was not used. In contrast, non-dieters recalled the same amount of information regardless of whether the labels were TLC color coded. Additionally, dieters accurately recalled more information from the TLC labels than non-dieters.

Trudel et al. (2015) concluded that dieters were more affected by the TLC food labeling. Their first three experiments were laboratory experiments, but they also wanted to see if the effects could be observed outside the laboratory. Experiment 4 was a *field study* where they attempted to replicate their laboratory results with grocery store shoppers. Shoppers at the entrance to a grocery store received a description and either mostly green or mostly red TLC nutrition labels of chocolate candy. They then were offered to sample as many chocolates as they wished from a bowl of 25 chocolates. Next, they rated their perceptions of health of the chocolates on the 9-point scale described above. Trudel et al. found that dieters rated the chocolates as equally healthy, regardless of red or green TLC labels. This effect was slightly different from the results of the laboratory experiments in which dieters showed lower health ratings of food with TLC colored labels than no colored labels. Non-dieters, however, rated the green-labeled chocolates as healthier than the red-labeled chocolates, consistent with the laboratory results. Furthermore, non-dieters took more chocolates to eat than dieters, and especially if the labels were green. Incidentally, the authors also used a moderated *mediation analysis* (see Chapter 9) to demonstrate that shoppers' health evaluations predicted their consumption of chocolate and this depended on whether shoppers were dieters or non-dieters.

In summary, the Trudel et al. (2015) study is an excellent illustration of how applied research can solve real problems while contributing to our knowledge of fundamental psychological phenomena. The authors concluded from their experiments TLC labels influenced food preferences and food consumption and such influences differed between dieters and non-dieters. Non-dieters used the TLC labels as a more explicit guide (stop-go decision making) for evaluating the health quality of foods. Dieters used the TLC labels to more deeply process and remember more nutrition information, which in turn was related to lower health ratings and more self-regulatory control of food consumption.

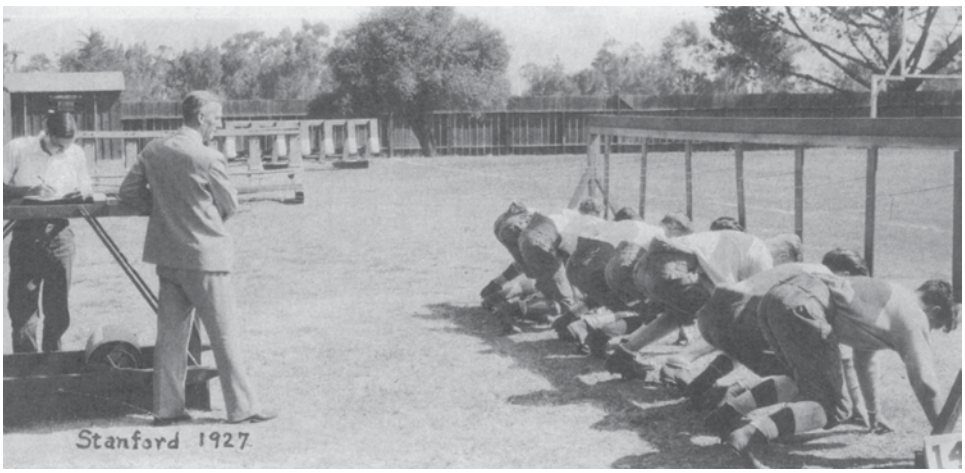
Applied Psychology in Historical Context

Because psychology in the United States developed in an academic setting, you might think research in psychology traditionally has been biased toward basic research. Not so. From the time psychology emerged as a new discipline in the late 19th century, psychologists in the United States have been interested in applied research and in applying the results of their basic research. For one thing, institutional pressures in the early 20th century forced psychologists to show how their work could improve society. In order to get a sufficient piece of the academic funding pie at

a time when psychology laboratories were brand new entities, psychologists had to show the ideas deriving from their research could be put to good use.

Psychologists trained as researchers focused on extending knowledge, but they often found themselves trying to apply basic research methods to solve problems in areas such as education, mental health, child rearing, and, in the case of Walter Miles, sports. Miles was director of the psychology laboratory at Stanford University in the 1920s. Although devoted to basic research throughout most of his career, he nonetheless found himself on the football team's practice field in 1927, as shown in Figure 11.1 (that's Miles in the suit). Stanford's legendary football coach, "Pop" Warner, was known as an innovator, open to anything that might improve his team's performance. Enter Miles, who built what he called a "multiple chronograph" as a way of simultaneously testing the reaction time of seven football players, an offensive line (Miles, 1928). On a signal that dropped seven golf balls onto a cylinder rotating at a constant speed (one ball per player), the players would charge forward, pushing a board that pulled a string that released a second set of golf balls onto the drum. The balls left marks on the cylinder and, knowing the speed of the rotation and the distance between the marks, Miles was able to calculate the players' reaction times. Miles published several studies with his multiple chronograph (e.g., Miles, 1931) and demonstrated its usefulness for identifying the factors that affected a football player's "charging time," but the apparatus never enjoyed widespread use (Baugh & Benjamin, 2006). Nonetheless, it is a good example of an experimental psychologist using a basic laboratory tool—reaction time in this case—to deal with a concrete problem: how to improve the efficiency of Stanford's football team.

Walter Miles made just an occasional foray into applied psychology, devoting most of his life to basic experimental research. Other psychologists, however, while trained as experimental psychologists, turned applied psychology into a career. A prime example is Harry Hollingworth, who entered applied psychology simply to make enough money for his talented wife, Leta, to attend graduate school. The result was a classic study on drug effects, financed by the Coca-Cola Company, whose product had been seized in a raid in Tennessee in 1909 on the grounds that it contained what was considered to be a dangerous drug. Box 11.1 elaborates this fascinating story and describes an early example of a nicely designed double-blind drug effect experiment.



The Drs. Nicholas and Dorothy Cummings Center for the History of Psychology, The University of Akron.

FIGURE 11.1
Simultaneously testing the reaction times of Stanford football players, circa 1927 (from Archives of the History of American Psychology, University of Akron, Akron, Ohio).

BOX 11.1 CLASSIC STUDIES—The Hollingworth's, Applied Psychology, and Coca-Cola

In 1911, the Coca-Cola Company was in some danger of having its trademark drink removed from the market or, at the very least, having one of its main ingredients removed from the formula. Under the federal Pure Food and Drug Act, which had been passed 5 years earlier during a time of progressive reform, Coca-Cola stood accused of adding a dangerous chemical to its drink: caffeine. It was said to be addictive (and they sell it to children!), and its stimulant properties were said to mask the need we all have for rest when we become fatigued. In 1909, a shipment of Coke syrup was seized by federal agents in Tennessee. Two years later, Coca-Cola found itself in court, defending its product. Enter Harry and Leta Hollingworth and a research story documented by Benjamin, Rogers, and Rosenbaum (1991).

In 1911, Harry Hollingworth was a young professor of psychology at Barnard College in New York City, anticipating a typical academic career of teaching, doing research, and avoiding committee work. His wife, Leta, also aimed for a professional career as a psychologist. They had married a few years earlier, and the plan was for Leta to teach school in New York while Harry finished graduate studies and began his career. Then Leta would go to graduate school. Unfortunately, Leta soon discovered one of the realities of being a married woman in early 20th century New York: The city's school board did not allow married women to teach (being married was assumed to be a woman's career). Financial difficulties immediately beset the young couple and, when the Coca-Cola Company offered Harry money to examine the cognitive and behavioral effects of caffeine, financial necessity prompted him to agree. To his credit, Hollingworth insisted (and Coca-Cola agreed) on being allowed to publish his results, whether or not they were favorable to the company.

Harry and Leta collaborated on the design for the studies they completed, an elaborate series of experiments that lasted more than a month. With Coca-Cola money, they rented a large apartment in which to conduct the research, with the daily data-collection sessions under Leta's supervision. Five rooms were set up as separate labs, with graduate students serving as experimenters. A variety of mental and physical tests were used, ranging from reaction time to motor coordination. Sixteen subjects were tested. Methodologically, the Hollingworths put into practice several of the concepts you have been studying in your methods course.

- They used *counterbalancing*.

With $N = 16$ and a month's worth of research, you can easily guess that each subject was tested many times. As with any repeated measures design, order effects were controlled through counterbalancing. For example, in one of the studies, participants rotated among five rooms in the apartment, completing a series of tests in each room. The order in which participants were tested in the rooms was randomized, in essence a partial counterbalancing procedure.

- They used a *placebo* control.

Participants were tested after taking pills containing either caffeine or a sugar substance. One set of studies included four groups, a placebo control, and three caffeine groups, each with a different dosage. Hence, the Hollingworths were able to examine not just caffeine effects but also dosage effects.

- They used a *double-blind* procedure.

Subjects did not know if they were receiving caffeine or a placebo, and the experimenters doing the tests in each room did not know if their subjects had taken caffeine or the placebo (Leta, who was in overall command of the testing, knew.)

And the results? Complex, considering the large of number tests used, the dosages employed, a fair amount of individual variation in performance, and the absence of sophisticated inferential statistical techniques (remember from Box 8.3 that nobody was doing ANOVAs before the mid-1930s). In general, no adverse effects of caffeine were found, except that larger doses, if taken near the end of the day, caused some subjects to have difficulty with sleep. Writing several years later in the textbook *Applied Psychology* (Hollingworth & Poffenberger, 1917), Harry wrote that the "widespread consumption of caffeinic beverages . . . seems to be justified by the results of these experiments" (p. 181). As for the trial, Harry testified on behalf of the company, arguing there was no scientific basis for banning caffeine in Coca-Cola. The case against Coke was eventually dismissed (on grounds other than caffeine's effects). One final outcome of the study was that it indeed paid for Leta's graduate studies, with enough money left over for a summer-long European trip. Leta Hollingworth eventually became a pioneer in the study and education of gifted children, probably better known than her husband.

Psychologists at the beginning of the 21st century are as interested in application as were their predecessors at the beginning of the 20th century. That is, they design and carry out studies to help create solutions to real-world problems while at the same time contributing to the basic core knowledge of psychology. However, applied research projects encounter several difficulties not usually found in the laboratory.

Design Problems in Applied Research

From what you have already learned in Chapters 2, 5, and 6, you should be able to anticipate most of the problems encountered in applied research, which include:

- *Ethical dilemmas* (Chapter 2). A study conducted outside of the laboratory may create problems relating to informed consent and privacy. Also, proper debriefing is not always possible. Research done in an industrial or corporate setting may include an element of perceived coercion if employees believe their job status depends on whether they volunteer to participate in a study (see Box 11.3 at the end of this chapter for more on ethics and applied research).

- *A trade-off between internal and external validity* (Chapter 5). Because research in applied psychology often takes place in the field, the researcher can lose methodological control over the variables operating in the study. Hence, the danger of possible confounding can reduce the study's internal validity. On the other hand, external (and specifically, ecological) validity is usually high in applied research because the setting more closely resembles real-life situations, and the problems addressed by applied research are everyday problems.

- *Problems unique to between-subjects designs* (Chapter 6). In applied research, it is often impossible to use random assignment to form equivalent groups. Therefore, the studies often use ex post facto designs and must therefore compare nonequivalent groups. This, of course, introduces the possibility of reducing internal validity by subject selection problems or interactions between selection and other threats such as maturation or history. When matching is used to achieve a degree of equivalence among groups of subjects, regression problems can occur, as will be elaborated in a few pages.

- *Problems unique to within-subjects designs* (Chapter 6). It is not always possible to counterbalance properly in applied studies using within-subjects factors. Hence, the studies may have uncontrolled order effects. Also, attrition can be a problem for studies that extend over a long period of time.

Before going much farther in this chapter, you might wish to look back at the appropriate sections of Chapters 2, 5, and 6 and review the ideas just mentioned. You also might review the section in Chapter 5 about the kinds of conclusions that can be drawn from manipulated variables and subject variables.

Self Test 11.1

1. Applied research is said to have a dual function. Explain.
2. Use the example of Miles and the Stanford football team to show how basic research and applied research can intersect.
3. How does applied research fare in terms of internal and external validity?

Quasi-Experimental Designs

Strictly speaking, and with Woodworth's (1938) definitions in mind, so-called true experimental studies include manipulated independent variables and equivalent groups formed by either straight random assignment or matching followed by random assignment. If subjects cannot be assigned randomly, however, the design is called a **quasi-experimental design**. Although it might seem that quasi-experiments are therefore lower in status than "true" experiments, it is important to stress that quasi-experiments have great value in applied research. They do allow for a degree of control, they serve a researcher's goals when ethical or practical problems make random assignment impossible, and they often produce results with clear benefits for people's lives. Thus far, we have seen several examples of designs that could be considered quasi-experimental:

- Single-factor ex post facto designs, with two or more levels
- Ex post facto factorial designs
- P x E factorial designs (the *P* variable, anyway)
- All of the correlational research

In this chapter, we will consider two specific designs typically found in texts on quasi-experimental designs (e.g., Cook & Campbell, 1979): *nonequivalent control group designs* and *interrupted time series designs*. Other quasi-experimental designs exist (e.g., regression discontinuity design), but these two are the most frequently encountered.

Nonequivalent Control Group Designs

In this type of study, the purpose is to evaluate the effectiveness of some treatment program. Those in the program are compared with those in a control group who aren't treated. This design is used when random assignment is not possible, so in addition to the levels of the independent variable, the members of the control group differ in some other way(s) from those in the treatment group—that is, the groups are not equivalent at the outset of the study. You will recognize this as a specific example of ex post facto design in Chapters 7 and 8, a type of design comparing nonequivalent groups, often selected with reference to a subject variable such as age, gender, or some personality characteristic. In the case of the quasi-experimental **nonequivalent control group design**, the groups are not equal at the start of the study; *in addition*, they experience different events in the study itself. Hence, there is a built-in confound that can complicate the interpretation of these studies. Nonetheless, these designs effectively evaluate treatment programs when random assignment is impossible.

Following the scheme first outlined by Campbell and Stanley (1963), the nonequivalent control group design can be symbolized like this:

Experimental group:	O_1	T	O_2
Nonequivalent control group:	O_1		O_2

where O_1 and O_2 refer to pretest and posttest observations or measures, respectively, and T refers to the treatment program being evaluated. Because the groups might differ on the pretest, the important comparison between the groups is not simply a test for differences on the posttest, but a comparison of the amounts of change from pre- to posttest in the two groups. Hence, the statistical comparison is typically between the *change scores* (the difference between O_1 and O_2) for each group. Alternatively, techniques are available that adjust posttest scores based on the pretests. Let's make this a bit more concrete.

Suppose the management of an electric fry pan company wants to institute a new flextime work schedule. Workers will continue to work 40 hours per week, but the new schedule allows them to begin and end each day at different times or to put all of their hours into 4 days if they wish to have a 3-day weekend. Management hopes this will increase productivity by improving morale and designs a quasi-experiment to see if it does. The company owns two plants in two very similar U.S. cities, one just outside of Pittsburgh and the other near Cleveland. Through a coin toss, the managers decide to make Pittsburgh's plant the experimental group and Cleveland's plant the nonequivalent control group. Thus, the study is quasi-experimental for the obvious reason that workers cannot be randomly assigned to the two plants (imagine the moving costs, legal fees over union grievances, etc.). The independent variable is whether or not flextime is present, and the dependent variable is some measure of productivity. Let's suppose the final design looks like this:

Pittsburgh plant:	pretest:	average productivity for 1 month prior to instituting flextime
	treatment:	flextime instituted for 6 months
	posttest:	average productivity during the sixth full month of flextime
Cleveland plant:	pretest:	average productivity for 1 month prior to instituting flextime in Pittsburgh
	treatment:	none
	posttest:	average productivity during the sixth full month that flextime is in effect in the Pittsburgh plant

Outcomes

Figure 11.2 shows you four outcomes of this quasi-experiment. All the graphs show the same amount of positive change in productivity for the Pittsburgh plant. The question is whether the change was due to program effectiveness or to some other factor(s). Before reading on, try to determine which of the graphs provides the strongest evidence that introducing flextime increased productivity. Also, refer to the section in Chapter 5 that described threats to internal validity, and try to identify the threats that make it difficult to interpret the other outcomes.

You probably found it fairly easy to conclude that in Figure 11.2a something besides the flextime produced the apparent improvement. This graph makes the importance of control groups obvious, even if it has to be a nonequivalent control group. Yes, Pittsburgh's productivity increased, but the same amount of change happened in Cleveland. Therefore, the Pittsburgh increase cannot be attributed to the program, but it could have been due to several of the threats to internal validity you've studied. *History* and *maturation* are good possibilities. Perhaps a national election intervened between pre- and posttest, and workers everywhere felt more optimistic, leading to increased productivity. Perhaps workers just showed improvement with increased experience on the job.

Figure 11.2b suggests that productivity in Cleveland was high throughout the study but that in Pittsburgh, productivity began at a very low level but improved due to the flextime program. However, there are two dangers here. For one thing, the Cleveland scores might reflect a *ceiling effect* (Chapter 5)—that is, their productivity level was so high to begin with that no further improvement could possibly be shown. If an increase could be seen (i.e., if scores on the *Y*-axis could go higher), you might see two parallel lines, as in Figure 11.2a. The second problem is that because Pittsburgh started so low, the increase there might be due to *regression to the mean* (Chapter 5) rather than a true effect. In other words, perhaps at the start of the study, productivity was very low for some reason, and it then returned to normal.

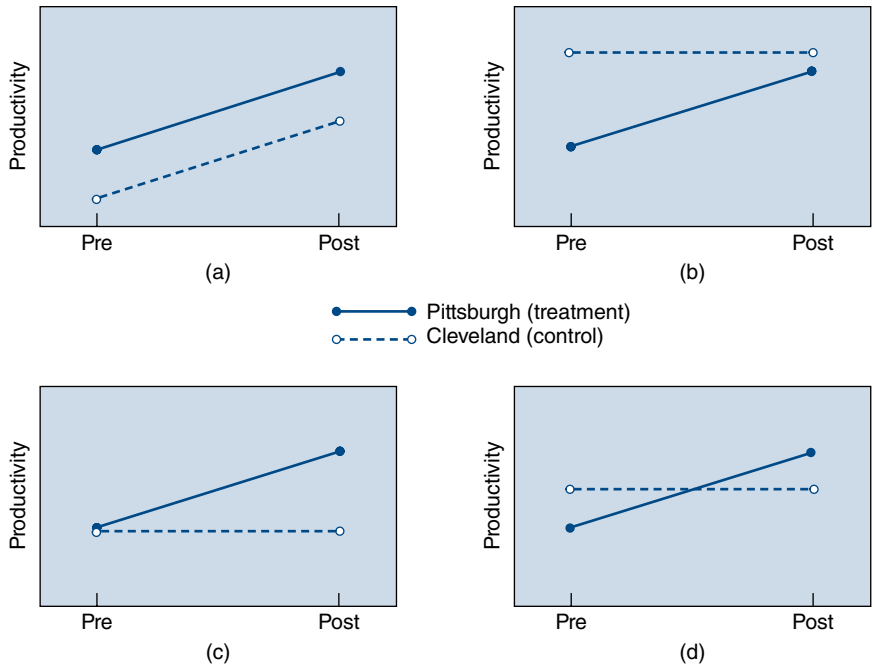


FIGURE 11.2
Hypothetical outcomes of a nonequivalent control group design.

Figure 11.2c seems at first glance to be the ideal outcome. Both groups start at the same level of productivity, but the group with the program (Pittsburgh) is the only one to improve. This may indeed be the case, and such an outcome generally makes applied researchers happy, but a problem can exist nonetheless. Because of the nonequivalent nature of the two groups, it is conceivable that subject selection could *interact* with some other influence—that is, history, maturation, or some other factor could affect the workers in one plant but not those in the other. For example, it's not hard to imagine a Selection x History problem here—some historical event affects the Pittsburgh plant but not the Cleveland plant. Perhaps the knowledge that they are participating in a study motivated the Pittsburgh workers (remember the *Hawthorne effect*?), while Cleveland workers were left in the dark. Perhaps between the pretest and the posttest, the Steelers won yet another Super Bowl, and because workers in Pittsburgh are such avid fans, their general feeling of well-being improved morale and therefore productivity. The Browns, on the other hand, who never win Super Bowls, would be less likely to inspire productivity boosts in the Cleveland plant.

The outcome in Figure 11.2d provides strong support for program effectiveness. Here, the treatment group (Pittsburgh) begins below the control group (Cleveland) yet surpasses the control group by the end of the study. Regression to the mean can be ruled out as causing the improvement for Pittsburgh because one would expect regression to raise the scores only to the level of the control group, not beyond it. Of course, selection problems and interactions between selection and other factors are difficult to exclude completely, but this type of crossover effect is considered good evidence of program effectiveness (Cook & Campbell, 1979).

Regression to the Mean and Matching

A special threat to the internal validity of nonequivalent control group designs occurs when there is an attempt to reduce the nonequivalency of the groups through a form of matching. Matching was first described in Chapter 6 as an alternative to random assignment under certain

circumstances, and it works rather well to create equivalent groups if the independent variable is a manipulated variable and participants can be randomly assigned to groups *after* being paired on some matching variable (see Chapter 6 to review the matching procedure). However, it can be a problem in nonequivalent control group designs when the two groups are sampled from populations that differ on the factor being used as the matching variable. If this occurs, then using a matching procedure can enhance the influence of the regression to the mean problem and even make it appear that a successful program has failed. Let's consider a hypothetical example.

Suppose you are developing a program to improve the reading skills of disadvantaged youth in a particular city. You advertise for volunteers to participate in an innovative reading program and select those most in need (i.e., those whose reading scores are, on average, very low). To create a control group that controls for socioeconomic class, you recruit additional volunteers from similar neighborhoods in other cities. Your main concern is equating your experimental and control groups for initial reading skill, so you decide to match the two groups on this variable. You administer a reading skills pretest to the volunteers in your target neighborhood and to the potential control group participants, and use the results to form two groups with the same average score—that is, the matching variable is the reading skills score. Let's say the test has a range from 0 to 100. You decide to select children for the two groups so the average score is 25 for both groups. The treatment group then gets the program and the control group does not; the design is a typical nonequivalent control group design:

Experimental group:	pretest	reading program	posttest
Control group:	pretest	—	posttest

Naturally, you're excited about the prospects of this study because you believe the reading program is unique and will help a lot of children. Hence, you're shocked when these reading scores occur:

Experimental group:	pre = 25	reading program	post = 25
Control group:	pre = 25	—	post = 29

Not only did the program not seem to work but also it appears it even hindered the development of reading skills—the control group apparently improved! What happened?

A strong possibility here is that regression to the mean resulting from the matching procedure overwhelmed any possible treatment effect. Remember that the experimental group was formed from those with the greatest need for the program because their skills were so poor. If the reading pretest could have been given to all children who fall into this category (i.e., the population called "very poor readers"), the average score might be quite low—let's say 17. When using the matching procedure, however, you were forced to select children who scored much higher than the mean score from this population of poor readers. Presumably, at least some of the children in this group scored higher than they normally would have on the pretest because no test is perfectly reliable—some degree of measurement error is likely to occur. Therefore, on a posttest, many of these children will score lower (i.e., move back to the mean of 17) simply due to regression to the mean. Let's suppose the program truly was effective and would add an average of 4 points to the reading score. However, if the average regression effect was a loss of 4 points, the effects would cancel each other out, and this would account for the apparent lack of change from pre- to posttest.

For participants in the control group, just the opposite might have occurred. Suppose their population mean score was higher than 25 (35, perhaps). Maybe they were reasonably poor readers to begin with, but not as bad as the experimental group (i.e., they were from a population called "relatively poor readers"). Selecting participants who scored lower than their population

mean, in order to produce pretest scores to match those of the experimental group (i.e., 25), could result in a regression effect producing higher posttest scores. For these children, the posttest score would result from the same regression to the mean found in the experimental group, although this time the regression effect would yield an increased score.

Figure 11.3 shows the problem in visual form. Regression and program improvements cancel each other out in the experimental group, while in the control group, regression is the only factor operating, and it pushes the scores toward the higher end. In sum, the reading program might actually have been a good idea, but the matching procedure caused a regression effect that masked its effectiveness.[2]

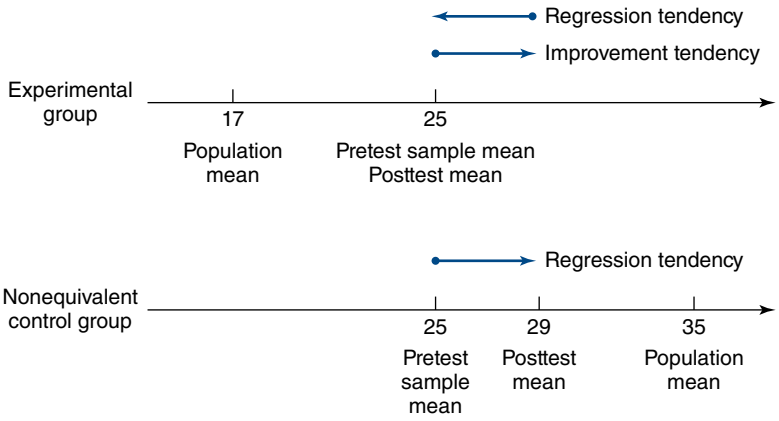


FIGURE 11.3
Hypothetical influences of regression to the mean when matching is used with nonequivalent groups, in an attempt to create equivalent groups.

This type of regression artifact apparently occurred during the first large-scale attempt to evaluate the effectiveness of Head Start, one of the cornerstone programs of President Lyndon Johnson's Great Society initiative in the 1960s (Campbell & Erlebacher, 1970). The program originated in 1965 as an ambitious attempt to give underprivileged preschool children a "head start" on school by teaching them school-related skills and getting their parents involved in the process. By 1990, about 11 million children had participated, and Head Start is now recognized as perhaps the most successful social program ever run by the federal government (Horn, 1990). Yet in the early 1970s, it was under attack for its "failure" to produce lasting effects, largely on the basis of what has come to be known as the Westinghouse study (because the study was funded by a grant to the Westinghouse Learning Corporation and Ohio University), conducted by Victor Cicirelli and his colleagues (1969).

The Westinghouse study documented what it called "fade-out effects"; early gains by children in Head Start programs seemed to fade away by the third grade. The implication, of course, was that perhaps federal dollars were being wasted on ineffective social programs, a point made by President Nixon in an address to Congress in which he explicitly referred to the Westinghouse study. Consequently, funding for Head Start came under attack during the Nixon years. At the same time, the basis for the criticism, the Westinghouse study, was being assaulted by social scientists.

Because Head Start was well under way when the Westinghouse evaluation project began, children couldn't be randomly assigned to treatment and control groups. Instead, the Westinghouse

[2] Although the practical and ethical realities of applied research in the field might prevent it, a better procedure would be to give a large group of children the reading readiness test, match them on the test, and randomly assign them to a reading program group and a control group.

group selected a group of Head Start children and matched them for cognitive achievement with children who hadn't been through the program. However, in order to match the groups on cognitive achievement, Head Start children selected for the study were those scoring well above the mean for their overall group, and control children were those scoring well below the mean for their group. This is precisely the situation described in the hypothetical case of a program to improve reading skills. Hence, the Head Start group's apparent failure to show improvement in the third grade was at least partially the result of a regression artifact caused by the matching procedure, according to Campbell and Erlebacher (1970).

In fairness to the Westinghouse group, it should be pointed out they would have disagreed vehemently with politicians who wished to cut the program. Cicirelli (1984) insisted that the study "*did not conclude that Head Start was a failure*" (p. 915; italics in the original), that more research was necessary and that "vigorous and intensive approaches to expanding and enriching the program" (p. 916) should be undertaken. Cicirelli (1993) later pointed out a key recommendation of the Westinghouse study was "not to eliminate Head Start but to try harder to make it work, based on encouraging findings from full-year programs [as opposed to summer-only programs]" (p. 32).

Nonequivalent control group designs do not always produce the type of controversy that engulfed the Westinghouse study. Consider the following two research examples: one is an attempt to increase the physical activity in children during school vacations, and the second is a study of the psychological aftershocks of an earthquake. Although most nonequivalent control group designs use pretest–posttest designs, the second study shows that nonequivalent control group designs do not always use pretests.

Research Example 34—A Nonequivalent Control Group Design

Many health benefits are associated with physical activity, particularly in children. In its FITT plan for physical activity, where FITT stands for frequency, intensity, time, and type, the American Academy of Pediatrics (AAP) recommends daily physical activity that is at least moderately vigorous in intensity (healthychildren.org). However, more access and use of screen-based activities (video games and television) is associated with less physical activity and more sedentary time in children. While there are specific physical activity plans (e.g., FITT) designed to help families increase children's physical activity levels, there are also barriers to access to physical spaces in which children can play. One barrier may be unsafe city streets due to car traffic. In the city of Ghent, Belgium, researchers with the consent of the city council opened up some city streets as car-free spaces where children could play safely.

D'Haese, Van Dyck, Bourdeaudhuij, Deforche, and Cardon (2015) used a nonequivalent control groups design to test whether the opening of "Play Streets" would increase physical activity and decrease sedentary activity during the summer months when children were not in school. Play Streets in Belgium are streets reserved for children's safe play during certain days and times, as determined by city councils. Usually, car traffic is prohibited and the streets are opened during the summer months when children are not in school. D'Haese et al. *operationally defined* their independent variable as neighborhoods where Play Streets had occurred versus not. Thus, the experimental group included children who lived in neighborhoods with Play Streets and the control group included children who lived in comparable neighborhoods but without Play Streets. Importantly, Play Streets occurred on only some summer days in those designated neighborhoods, so the researchers could use a pretest–posttest design to measure physical activity before Play Streets and after Play Streets in the experimental condition. The design looked like this:

Experimental group:	pretest	Play Streets	posttest
Nonequivalent control group:	pretest	No Play Streets	posttest

Recall that in nonequivalent control group designs, the experimental and control group are not equal at the start, but attempts are made to make the groups as similar as possible. This was the case in the Play Streets study. The two groups were unequal in that they were different city neighborhoods, but D'Haese et al. ensured that the neighborhoods in both groups were similar in terms of walkability of the neighborhoods and the annual household incomes of its residents.

To measure children's physical activity, children wore accelerometers all day including during the 5 hours when Play Streets were open for public use. The accelerometers enabled researchers to record how long children engaged in moderate-to-vigorous physical activity (MVPA). D'Haese et al.'s (2015) results are displayed in Figure 11.4. As you can see, both children in both types of neighborhoods engaged in equivalent levels of MVPA per day before the implementation of Play Streets. After Play Streets, children in the experimental group showed an increase in physical activity, whereas the children in the control group became less physically active.

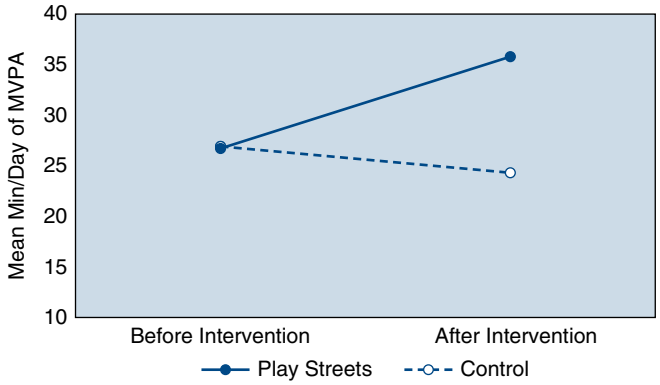


FIGURE 11.4
Changes in (a) physical activity and (b) sedentary activity after implementation of a Play Streets program. Activity recorded during the hours of the Play Streets program. (from D'Haese et al., 2015).

Interestingly, the effects carried over into the entire day. During days when Play Streets were implemented (or not in the case of the control group), children in the Play Streets condition increased their total MVPA from 55 min/day to 67 min/day, whereas children in the control condition decreased their total MVPA from 57 min/day to 53 min/day. Given recommendations that children should engage in physical activity for about 60 min/day, it is evident that use of a program like Play Streets can help provide safe, fun places for children to increase their physical activity to recommended levels.

In addition to using a nonequivalent control group design, D'Haese et al. (2015) also used a *survey method* in which they asked parents' opinions about Play Streets. They asked parents in experimental and control conditions to rate several items on a Likert scale in terms of their level of agreement with various statements. Approximately 60% of parents whose children had access to Play Streets either agreed or strongly agreed that they had the impression that their children played more outside, which was confirmed by children's actual physical activity levels in the experimental group. For parents who were part of the control group (no Play Streets), 76% of parents agreed or strongly agreed that if they had Play Streets that their children would have more social contact with each other. The results of a comparable item for the Play Streets condition showed that 78% of parents reportedly they believed that their children had many friends in the Play Street. The authors concluded that in addition to the health benefits of increased physical activity, social interactions among children can be enhanced with the use of Play Streets. Finally, programs like Play Streets can be very low cost, particularly for low income communities where access to public parks or playgrounds may be limited. In this case, while costs may be low, the

benefits to children may be quite high. Later in this chapter, we will explore in more depth *cost-effectiveness analysis* when it comes to a research process called *program evaluation*.

Research Example 35—A Nonequivalent Control Group Design Without Pretests

Nonequivalent control group designs typically include pretests but that is not always the case. Sometimes, these designs occur when an unforeseen opportunity for research makes pretesting impossible. One such event was the 1989 San Francisco earthquake. To James Wood and Richard Bootzin of the University of Arizona, the event suggested an idea for a study about nightmares, a topic already of interest to them (Wood & Bootzin, 1990). Along with colleagues from Stanford University (located near the quake's epicenter), they quickly designed a study to see if the experience of such a traumatic event would affect dream content in general and nightmares in particular (Wood, Bootzin, Rosenhan, Nolen-Hoeksema, & Jourden, 1992). By necessity, they used a nonequivalent control group design. As is generally the case with this design, the groups were nonequivalent to begin with (students from two states); in addition, one group had one type of experience (direct exposure to the earthquake), while the second group had a different experience (no direct exposure).

The experimental group consisted of students from Stanford University and San Jose State University who experienced the earthquake. Nonequivalent controls were college students recruited from the University of Arizona. They did not experience the quake, of course, but they were well aware of it through the extensive media accounts. Shortly after the earthquake event (about a week), all participants began keeping a dream log, which was then analyzed for nightmare content and frequency. Wood et al. (1992) were careful to provide a clear *operational definition* of a nightmare ("frightening dreams with visual content and an elaborated story," p. 220) and to differentiate nightmare from night terrors ("awakening during the night with feelings of intense fear or terror but no memory of a dream," p. 220), instructing subjects to report the former only.

The results were intriguing. Over the 3 weeks of the study, 40% of San Jose students and 37% of the Stanford students reported having at least one earthquake nightmare, while only 5% of the control students at Arizona did (Wood et al., 1992). Of the total number of nightmares experienced by the experimental groups, roughly one-fourth were about earthquakes (27% for San Jose, 28% for Stanford), but virtually, none of the control group's nightmares were about quakes (3% for Arizona). Furthermore, the frequency of nightmares correlated significantly with how anxious participants reported they were during the time of the earthquake.

Well aware of the interpretation problems that accompany quasi-experimental studies, Wood et al. (1992) recognized the dangers inherent in comparing nonequivalent groups. For instance, lacking any pretest (pre-quake) information about nightmare frequency for their participants, they couldn't "rule out the possibility that California residents have more nightmares about earthquakes than do Arizona residents even when no earthquake has recently occurred" (p. 222). If one lives in California, perhaps earthquake nightmares are a normal occurrence. However, relying partly on their general expertise in the area of nightmare research, and partly on other survey data about nightmares (estimates from subjects of pre-earthquake nightmares), the authors argued that the nightmare frequency was exceptionally high in the California group and likely the result of their recent traumatic experience.

Interrupted Time Series Designs

If Wood and his colleagues (1992) could have foreseen San Francisco's earthquake, they might have started collecting nightmare data from their participants for several months leading up to the quake and then for several months after the quake. That would have enabled them to determine (a) if the quake truly increased nightmare experiences for the participants in the quake zone, and (b) if the nightmare frequency peaked shortly after the quake and then returned to baseline. Of course, not even talented seismologists can predict earthquakes, so Wood and his coworkers

did the best they could and designed their nonequivalent control group study. If they had been able to take measures for an extended period before and after the event expected to influence behavior, their study would have been called an **interrupted time series design**.

Using the system in Campbell and Stanley (1963) again, the basic time series study can be symbolized like this:

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

where all of the O's represent measured observations of behavior taken before and after T, which is the point at which some treatment program is introduced or some event (e.g., an earthquake) occurs. T is the interruption in the interrupted time series. Of course, the number of measures taken before and after T will vary from study to study and are not limited to five each. It is also not necessary that the number of pre-interruption and post-interruption points be the same. As a general rule, the more data points, the better, and some experts (e.g., Orwin, 1997) recommend at least *50* pre-interruption data points.

Outcomes

The main advantage of a time series design is that it allows the researcher to evaluate **trends**, which are relatively consistent patterns of events that occur with the passing of time. For example, suppose you were interested in seeing the effects of a 2-month antismoking campaign on the number of teenage smokers in a community. The program might include persuasion techniques, peer counseling, showing the teens a smoked-out lung or two, and so on. Assuming you had a good measure of the smoking behavior, you could take the measure a month before and a month after introducing the program and perhaps get the results in Figure 11.5.

Did the program work? There certainly is a reduction in smoking from pre- to posttest, but it's hard to evaluate it in the absence of a control group (i.e., using a nonequivalent control group design). Yet, even without a control group, it might be possible to evaluate the campaign more systematically if not one but several measures were taken both before and after the program was put in place. Consider the possible outcomes in Figure 11.6, which examines the effect of the antismoking campaign by measuring smoking behavior every month for a year before and a year after the program (the solid-line portion of the graphs duplicates Figure 11.5).

Figure 11.7a is a good illustration of how an interrupted time series can identify trends. In this case, the reduction that looked so good in Figure 11.5 is shown to be nothing more than part of a general trend toward reduced smoking among adolescents. This demonstrates an important feature of interrupted time series designs: They can serve to rule out (i.e., falsify) alternative explanations of an apparent change from pre- to posttest.

Two other outcomes that raise questions about the program's effectiveness are seen in Figure 11.6b and 11.6c. In Figure 11.6b, smoking behavior was fairly steady before the campaign

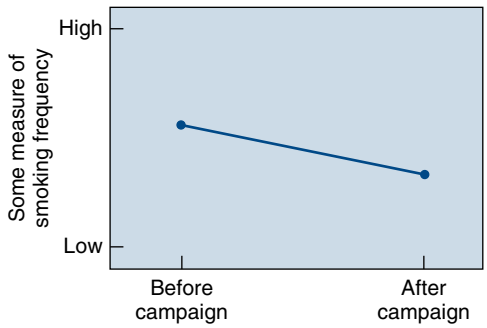


FIGURE 11.5
Incidence of smoking behavior just before and just after a hypothetical antismoking campaign.

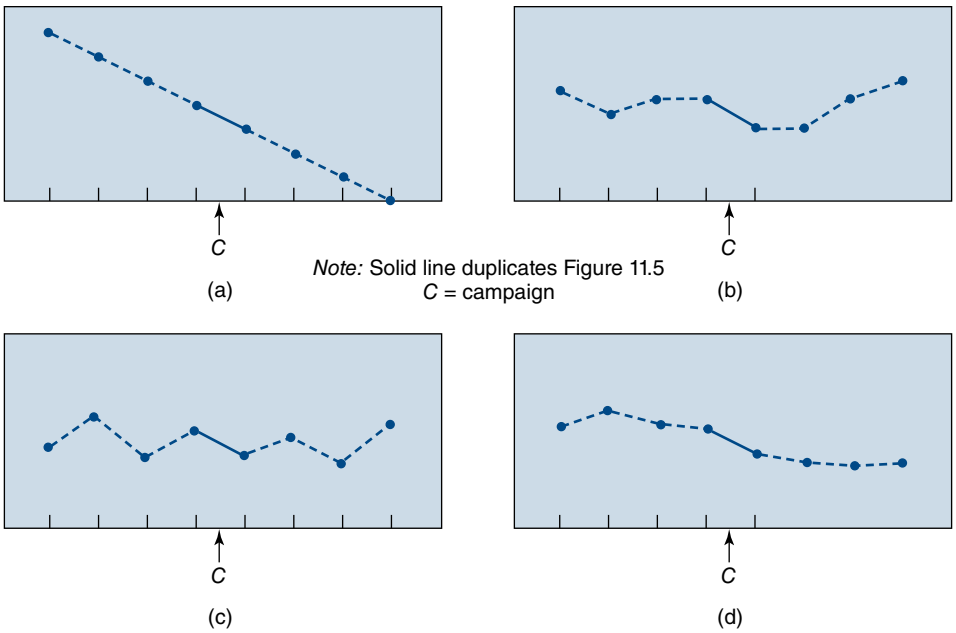


FIGURE 11.6
Hypothetical antismoking campaign evaluated with an interrupted time series design—several possible outcomes.

and then dropped but just briefly. In other words, if the antismoking program had any effect at all, it was short-lived. In Figure 11.6c, the decrease after the program was part of another general trend, this time a periodic fluctuation between higher and lower levels of smoking. The ideal outcome is shown in Figure 11.6d. Here the smoking behavior is at a steady and high rate before the program begins, drops after the antismoking program is put into effect, and remains low for some time afterward. Note also in Figure 11.6d that the relatively steady baseline prior to the campaign enables the researcher to rule out regression effects.

Research Example 36—An Interrupted Time Series Design

An actual example of an outcome like the one in Figure 11.6d can be found in a study of worker productivity completed at a unionized iron foundry by Wagner, Rubin, and Callahan (1988). They were interested in the effect of instituting an incentive plan in which workers were treated not as individuals but as members of small groups, each responsible for a production line. Productivity data were compiled for 4 years prior to introducing the incentive plan and 6 years afterward; there were 114 monthly data points. As you can see from their time series graph in Figure 11.7, productivity was fairly flat and not very impressive prior to the plan but increased steadily after the plan was implemented and stayed high for some time afterward.

This study also illustrates how those conducting interrupted time series designs try to deal with potential threats to internal validity. Figure 11.7 certainly appears to show the incentive plan worked wonders, but there is no control group comparison and the changes could have been influenced by other factors, including history, instrumentation, and even subject selection. Wagner et al. (1988) argued that history did not contribute to the change because they carefully examined as many events as they could in the period before and after the change and could find no reason to suspect that unusual occurrences led to the jump in productivity. In fact, events that might be expected to hurt productivity (e.g., a recession in the automobile industry, which affected sales of iron castings) didn't. The researchers also ruled out instrumentation, which

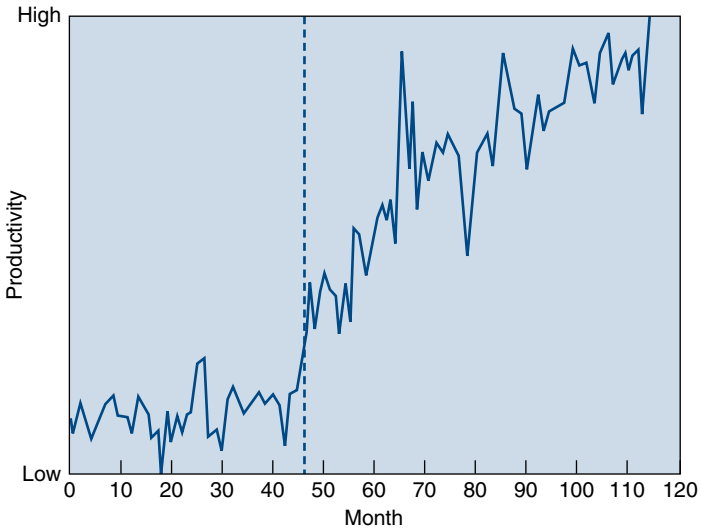


FIGURE 11.7
Interrupted time series design: effect of an incentive plan on worker productivity in an iron foundry (from Wagner et al., 1988).

could be a problem if the techniques for scoring and recording worker productivity changed over the years. It didn't. Third, although we normally think of subject selection as a potential confound only in studies with two or more nonequivalent groups, it can occur in a time series study if significant worker turnover occurred during the time of the new plan; the cohort of workers on site prior to the plan could be different in some important way from the group there after the plan went into effect. This didn't happen in Wagner et al.'s study though. In short, designs like this one, because they lack a control group, are susceptible to several threats to internal validity. These threats often can be ruled out, however, by systematically examining available information, and Wagner and his colleagues did just that.

Variations on the Basic Time Series Design

Sometimes, the conclusions from an interrupted time series design can be strengthened if some type of control comparison is made. One approach amounts to combining the best features of the nonequivalent control group design (a control group) and the interrupted time series design (long-term trend analysis). The design looks like this:

$$\begin{array}{lcl} O_1\ O_2\ O_3\ O_4\ O_5 & \mathbf{T} & O_6\ O_7\ O_8\ O_9\ O_{10} \\ O_1\ O_2\ O_3\ O_4\ O_5 & & O_6\ O_7\ O_8\ O_9\ O_{10} \end{array}$$

If you look ahead to Figure 11.9 in Box 11.2, you will see a classic example of this strategy, a study that evaluated a speeding crackdown in Connecticut by comparing fatal accident data from that state with data from similar states. Another example comes from the aftermath of the Oklahoma City bombing in 1995. This domestic terrorist attack, the bombing of a federal building, killed 168 persons, including 19 children, and injured more than 700. Nakonezny, Reddick, and Rodgers (2004) hypothesized that the resulting feelings of helplessness and insecurity would lead Oklahoma City residents to seek "the comfort and support of familial and marital bonds to restore a sense of structure and security" (p. 91). Focusing on divorce rates in Oklahoma City and in several comparison locations in the state for 10 years before the bombing and 5 years after it,

they discovered a significant decline in the Oklahoma City divorce rate for several years following the bombing. Like Figure 11.6b, however, the decline was not lasting; after 5 years, the divorce rate reverted to the normal statewide rate.

A second strategy for strengthening conclusions from a time series study is when a program can be introduced in different locations at different times, a design labeled an **interrupted time series with switching replications** by Cook and Campbell (1979), and operating like this:

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

With this procedure, the same treatment or program is put into place in two locations at two points in time. There is no control group, but the design provides the benefit of a built-in replication. If the outcome pattern in Location 2 matches that of Location 1, the researchers can be more confident about the generality of the phenomenon being studied. This happened in an unpublished study reported in Cook and Campbell (1979). It was completed in the late 1940s and early 1950s, when televisions were just starting to change our lives. A number of Illinois communities were given licenses for new TV stations, but in 1951, there was a freeze on new licenses that wasn't lifted until 1953. That gave researchers an opportunity to study the impact of new televisions on communities at two different times: in the late 1940s, just before the freeze, and right after 1953, with the freeze lifted. Hypothesizing that the new invention would reduce the amount of reading done, researchers studied library circulation data and found support for their concerns about reading. As TVs began infiltrating communities, library circulation dropped, and the pattern was virtually identical during the two times examined.

A third elaboration on an interrupted time series design, again in the absence of a control group, is to measure several *dependent* variables, some expected to be influenced by the interruption, others not expected to change. This was the strategy used in a study by Stolzenberg and D'Alessio (1997). They examined the effect of a California mandatory jail sentencing law, the "three strikes and you're out" policy, on crime rates. The essence of the policy is that jail sentences occur automatically once a person has been convicted of three serious crimes (felonies). Combining data from California's 10 largest cities, Stolzenberg and D'Alessio examined two types of crime rates (i.e., two dependent variables): felonies, supposedly reduced by mandatory sentencing, and misdemeanors (relatively minor crimes). Presumably, misdemeanors would not be affected by the three strikes law. Figure 11.8 shows the results, a good example of the advantages of a time series design. If you look at the curve for serious crimes right after the law was passed, it looks like there is a decline, especially when compared to the flat curve for the misdemeanors. If you look at the felony crime curve as a whole; however, it is clear that any reduction in serious crime is part of a trend occurring since around 1992, well before passage of the three strikes law. Overall, the researchers concluded the three strikes law had no discernible effect on serious crime.

Self Test 11.2

1. Why is it said that the nonequivalent control group design has a built-in confound?
2. If nonequivalent groups are used and the groups are matched on a pretest score, the results can be distorted by a _____ effect.
3. Time series designs sometimes include "switching replications." How does this design differ from the basic interrupted time series design?

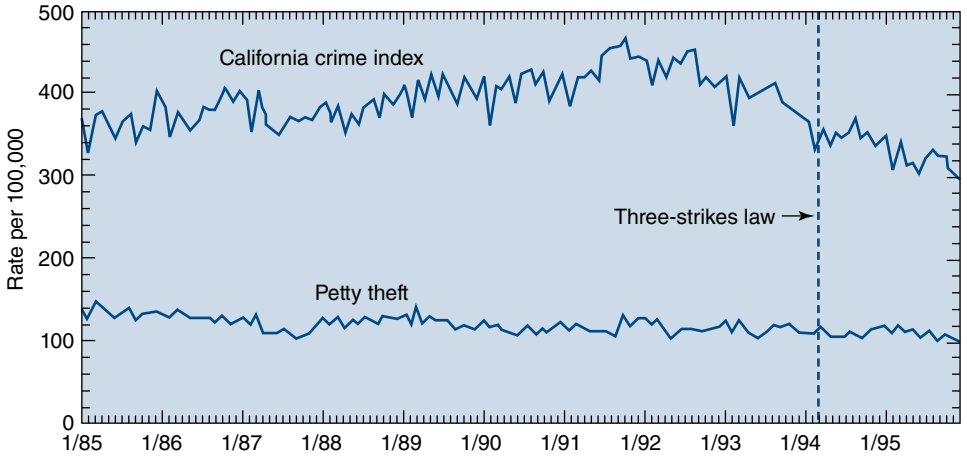


FIGURE 11.8
Interrupted time series using two different dependent measures; the effect of mandatory sentencing on crime rates (from Stolzenberg & D'Alessio, 1997).

Program Evaluation

Applied research that attempts to assess the effectiveness and value of public policy (e.g., California's three strikes law) or specially designed programs (e.g., Meals on Wheels) is sometimes given the name **program evaluation**. This research concept developed in the 1960s in response to the need to evaluate social programs like Head Start, but it is concerned with much more than answering the question "Did program X work?" More generally, program evaluation includes (a) procedures for determining if a need exists for a particular program and who would benefit if the program is implemented; (b) assessments of whether a program is being run according to plan and, if not, what changes can be made to facilitate its operation; (c) methods for evaluating program outcomes; and (d) cost analyses to determine if program benefits justify the funds expended. Let's consider each in turn. First, however, you should read Box 11.2, which highlights a paper by Donald Campbell (1969) that is always included at or near the top of lists of the "most important papers about the origins of program evaluation."

BOX 11.2 ORIGINS—Reforms as Experiments

A 1969 article by Donald Campbell entitled "Reforms as Experiments" is notable for three reasons. First, he argued forcefully that we should have an experimental attitude toward social reform. In the opening sentence, Campbell wrote:

> [W]e should be ready for an experimental approach to social reform, an approach in which we try out new programs designed to cure specific social problems, 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. (p. 409)

Second, Campbell's (1969) article helped originate and define the field of program evaluation, and it described several studies that have become classics. Perhaps the best-known example is his description of a study evaluating an effort to reduce speeding in Connecticut (Campbell & Ross, 1968). Following a year (1955) with a record number of traffic fatalities (324), Connecticut governor Abraham Ribicoff instituted a statewide crackdown on speeding, making the reasonable assumption that speeding and traffic fatalities were related. The following year, the number of deaths fell to 284. This statistic was sufficient for Ribicoff to declare that with

"the saving of 40 lives in 1956, a reduction of 12.3% from the 1955 . . . death toll, we can say that the program is definitely worthwhile" (quoted in Campbell, 1969, p. 412). Was it?

I hope you're saying to yourself that other interpretations of the drop are possible. For example, history could be involved; perhaps the weather was better in 1956. Even more likely is regression to the mean—324 is the perfect example of an extreme score that would normally be followed by regression to the mean. Indeed, Campbell argued that regression contributed to the Connecticut results, pointing out that "[r]egression artifacts are probably the most recurrent form of self-deception in the experimental social reform literature" (p. 414). Such effects frequently occur in these kinds of studies because interventions like a speeding crackdown often begin right after something especially bad has happened. Purely by chance alone, things are not likely to be quite as bad the following year.

Was regression to the mean all that was involved here? Probably not. By applying an interrupted time series design with a nonequivalent control (comparable states without a crackdown on speeding), Campbell concluded the crackdown probably did have some effect, even if it was not as dramatic as the governor believed. You can see the results for yourself in Figure 11.9.

The third reason the Campbell article is so important is that it gave researchers insight into the political realities of doing research on socially relevant issues. Politicians often propose programs they believe will be effective and, while they might say they're interested in a thorough evaluation, they tend not to be too appreciative of a negative evaluation. After all, by backing the program, they have a stake in its success and its continuance, especially if the program benefits the politician's home state or district. For this reason, politicians and the administrators hired to run programs seldom push for rigorous evaluation and are willing to settle for favorable research outcomes even if they come from flawed research design. For example, Governor Ribicoff was willing to settle for looking at nothing more than traffic fatalities immediately before and after the crackdown on speeding.

Campbell (1969) recommended an attitude change that would shift emphasis from the importance of a particular program to acknowledging the importance of the problem. This would lead politicians and administrators alike to think of programs as experimental attempts to solve the problem; different programs would be tried until one was found to work. As Campbell put it in the article's conclusion,

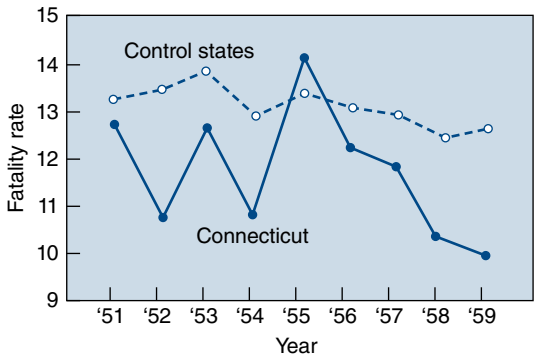


FIGURE 11.9
The Connecticut speeding crackdown, a classic example of an interrupted time series with a nonequivalent control (from Campbell, 1969).

> *Trapped administrators* have so committed themselves in advance to the efficacy of the reform that they cannot afford an honest evaluation. . . . *Experimental administrators* have justified the reform on the basis of the importance of the problem, not the certainty of their answer, and are committed to going on to other potential solutions if the first one tried fails. They are therefore not threatened by a hard-headed analysis of the reform. (p. 428; italics in the original)

Planning for Programs—Needs Analysis

An agency begins a program because administrators believe a need exists that would be met by the program. How is that need determined? Clearly, more is required than just an administrative decision that a program seems to make sense. An exercise program in a retirement community sounds reasonable, but if none of the residents will participate, time and money will be wasted. Before any project is planned in any detail, a needs assessment must be completed.

A **needs analysis** is a set of procedures for predicting whether a population of sufficient size exists that would benefit from the proposed program, whether the program could solve a clearly defined problem, and whether members of the population would actually use the program. Several

methods exist for estimating need, and it is important to rely on at least some of these techniques because it is easy to overestimate need. One reason for caution follows from the *availability heuristic*, first introduced in Chapter 1's discussion about ways of knowing. Events that grab headlines catch our attention and become more "available" to our memory. Because they come so readily to mind, we tend to overestimate how often they occur. All it takes is one or two highly publicized cases of children being abandoned by vacationing parents for a call to be made for new programs to fix this seemingly widespread problem. Also, a need for a new program can be overestimated by those in a position to benefit (i.e., keep their jobs) from the program's existence.

As outlined by Posavac and Carey (2010), there are several ways to identify the potential need for a program. These include:

- *Census data.* If your proposed program is aimed at the elderly, it's fairly obvious that its success will be minimal if few seniors live in the community. Census data (www.census.gov) can provide basic demographic information about the number of people fitting into various categories. Furthermore, the information is fine-grained enough for you to determine the number of single mothers under the age of 21, the number of people with various disabilities, the number of older adults below the poverty line, and so on.
- *Surveys of available resources.* There's no reason to begin a Meals on Wheels program if one already exists in the community and is functioning successfully. Thus, one obvious step in a needs analysis is to create an inventory of existing services that includes a description of who is providing the services, exactly which services are being provided, and an estimate of how many people are receiving the services.
- *Surveys of potential users.* A third needs analysis strategy is to administer a survey within the community, either to a broadly representative sample or to a target group identified by census data. Those participating could be asked whether they believe a particular program is needed.
- *Key informants, focus groups, and community forums.* A **key informant** is someone in the community who has a great deal of experience and specialized knowledge about the problem at hand that is otherwise unavailable to the researcher (Gilchrist & Williams, 1999). Such persons include community activists, clergy, people who serve on several social service agency boards, and so on. A **focus group** is a small group (typically 7-9 people) whose members respond to a set of open-ended questions about some topic, such as the need for a particular program (they might also be used to assess a program's progress or its outcome). Focus groups are often used as a follow-up to a community survey, but they also can be used to shape the questions that will appear in a survey. Finally, useful information can sometimes emerge from a **community forum**, an open meeting at which all members of a community affected by a potential program are invited to come and participate. Key informants, focus groups, and forums can all be helpful tools, but the researcher must be careful of weighing too heavily the arguments of an especially articulate (but perhaps nonrepresentative) informant, focus group member, or speaker at a forum.

The past few decades have seen an increased awareness in corporate America that profits are related to worker health. Consequently, companies frequently develop, implement, and evaluate programs for improving the health of their workers. The following study describes a large-scale example that began with a thorough analysis of need.

Research Example 37—Assessing Need in Program Evaluation

A needs analysis project was undertaken by the Du Pont Company prior to starting a program designed to promote healthy behaviors in the workplace (Bertera, 1990). The plan called for a series of changes that would affect over 110,000 employees at 100 worksites. The cost of putting such an ambitious plan into effect made it essential that need be demonstrated clearly.

The Du Pont needs assessment included an analysis of existing data on the frequency of various types of employee illnesses, employee causes of death, and the reasons for employee absence and disability over a 15-year period. One result was that employees making the least amount of money and performing the lowest ranking jobs were the highest on all major categories of illness. That finding told the evaluators that this particular subgroup of workers needed special attention.

Additional indicators that the health promotion program was needed came from a survey of existing company programs for enhancing health. The survey revealed a range of programs run by the medical staffs at the various plants, including programs on weight loss, smoking cessation, stress management, and the like. The programs tended to be one-time lectures or films, however, or counseling during company physical exams; there was minimal follow-up and no systematic evaluation of effectiveness. Employees were also surveyed to determine their knowledge of health-enhancing behaviors, their intention to change things like their eating habits, their self-assessments of whether their own behaviors were health-enhancing or not, and their preferences for a range of health programs.

On the basis of all of this information, Du Pont developed a comprehensive series of programs aimed at improving the health of its workers. These included training programs that went far beyond one-shot lectures, including creation of local employee Health Promotion Activity Committees, recognition and award programs for reaching certain health goals, and workplace climate changes (e.g., removing cigarette vending machines). Also, all workers completed a Health Risk Survey. The results generated a Health Risk Appraisal, which became part of the workers' personnel files and included an individualized plan for promoting healthy behaviors. On the basis of their needs assessment, the Du Pont Company instituted a company-wide program designed to improve workplace health, specifically targeting "smoking cessation, blood pressure control, and lipid control" (Bertera, 1990, p. 316).

Once the needs analysis is complete and the decision is made to proceed, details of the program can be planned and the program begun. Once the program is under way, the second type of evaluation activity begins.

Monitoring Programs—Formative Evaluation

Programs often extend over a considerable period. To wait for a year or so before doing a final evaluation of program effectiveness might be preferable from a methodological point of view, but what if it is clear in the first month that problems exist that could be corrected easily? That is, rather than waiting until the program's completion, why not carefully monitor the progress of the program while it is in progress? This monitoring is called a **formative evaluation**, and according to one analysis (Sechrest & Figueredo, 1993), it is the most common form of evaluation activity.

A formative evaluation can include several components. For one thing, it determines if the program is being implemented as planned. For example, suppose a local crisis hotline decides to develop a program aimed at the needs of young children who are home alone after school while their parents are working. One piece of the implementation plan is to make the hotline's phone number available and well known. A formative evaluation would determine whether the planned advertisements were placed online or in local newspapers at appropriate times and whether mass mailings of stickers with the hotline's number went out as planned. There's no point in trying to evaluate the effectiveness of the program if people don't even know about it.

Another general function of the formative evaluation is to provide data on how the program is being used. Borrowing a term from accounting, evaluators sometimes refer to this procedure as a **program audit**. Just as a corporate auditor might look for inconsistencies between the way inventories are supposed to be managed and the way they actually are managed, the program auditor examines whether the program as described in the agency's literature is the same as the program that is actually being implemented.

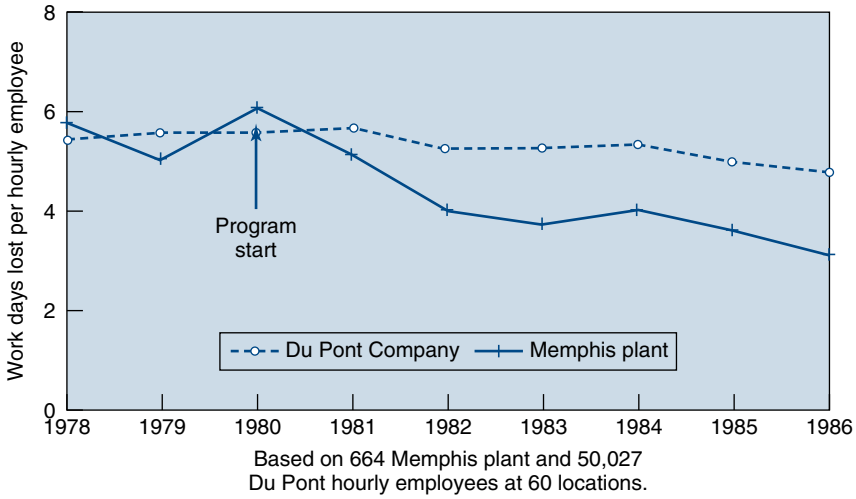


FIGURE 11.10
Effectiveness of a workplace health improvement program, evaluated via time series (from Bertera, 1990).

A final part of a formative evaluation can be a *pilot study* (Chapter 3). Program implementation and some preliminary outcomes can be assessed on a small scale before extending the program. This happened in the Du Pont study. A pilot program at one of the plants, which showed a significant decline in sick days after implementation of the health promotion program, encouraged program planners and led to an elaboration of the program at other sites (Bertera, 1990). As you can see from Figure 11.10, researchers used a time series design, with data collected over 8 years. (*Note*: By 1982, the results were clear enough that executives began expanding the program to other plants.)

Evaluating Outcomes—Summative Evaluation

Politically, formative evaluations are less threatening than **summative evaluations,** which are overall assessments of program effectiveness. Formative evaluation is aimed at program improvement and is less likely to call into question the program's very existence. Summative evaluation, on the other hand, can do just that. If the program isn't effective, why keep it, and, by extension, why continue to pay the program's director and staff? (See what we mean about "threatening?") As Sechrest and Figueredo (1993) stated:

> Summative evaluation and even the rationale for doing it call into question the very reasons for existence of the organizations involved. Formative evaluation, by contrast, simply responds to the question "How can we be better?" without strongly implying the question "How do [we] know [we] are any good at all?" (p. 661)

Despite the political difficulty, summative evaluations are the core of the evaluation process and are an essential feature of any program funded by the federal government. Any agency wishing to spend tax dollars to develop a program is obligated to show those dollars are being used effectively.

The actual process of performing summative evaluations involves applying some of the techniques you already know about, especially quasi-experimental designs. However, more rigorous experiments with random assignment are possible sometimes, especially when evaluating

a program that has more people desiring it than space available. In such a case, random assignment in the form of a lottery (random winners get the program; others wind up in a wait list control group) is not only methodologically sound, it is also the only fair procedure to use.

One problem that sometimes confronts the program evaluator is how to interpret a failure to find significant differences between experimental and control groups—that is, the statistical decision is "fail to reject the null hypothesis" Such an outcome is difficult to interpret, as you recall from the discussion in Chapter 4. It could be there just isn't any difference, yet there's always the possibility of a *Type II error* being committed (an effect is real, but your study failed to find it), especially if the measuring tools are not sensitive or reliable. The program might indeed have produced some small but important effect, but the analysis failed to discover it.

Although a finding of no difference can be difficult to interpret, most researchers believe that such a finding (especially if replicated) contributes important information for decision making, especially in applied research. For instance, someone advocating the continuation of a new program is obligated to show how the program is better than something already in existence. Yet, if differences between this new program and one already well established cannot be shown, then it might be wise to discontinue the new program, especially if it is more expensive to implement than the older one. A "fail to reject the null" decision also can help evaluate exaggerated claims made by advocates of a new program. A finding of no difference has important implications for decision making for reasons having to do with cost, and this brings us to the final type of program evaluation activity.

Weighing Costs—Cost-Effectiveness Analysis

Suppose a researcher is interested in the question of worker health and fitness and is comparing two health-enhancement programs. One includes opportunities for exercising on company time, educational seminars on stress management, and a smoking ban. The second plan is a more comprehensive (and more expensive) program of evaluating each worker and developing an individually tailored fitness program, along with financial incentives for achievements like reducing blood pressure and cholesterol levels. Both programs are implemented on a trial basis in two plants; a third plant is used as a control group. Hence, the design is a nonequivalent control group design with two experimental groups instead of just one. A summative evaluation finds no difference between the two experimental groups in terms of improved worker health, but both show improvements compared to the control group. In other words, both health programs work, but the cheap version works just as well as the expensive version. If two programs producing the same outcome differ in cost, why bother with the expensive one?

This corporate fitness example illustrates one type of **cost-effectiveness analysis**: monitoring the actual costs of a program and relating those costs to the effectiveness of the program's outcomes. If two programs with the same goal are equally effective but the first costs half as much as the second, then it is fairly obvious that the first program should be used. A second type of cost analysis takes place during the planning stages for a program. Estimating costs at the outset helps determine whether a program is feasible and provides a basis for the later comparison of projected costs and actual costs.

Estimating costs with reference to outcomes can be a complicated process, often requiring the expertise of a specialist in cost accounting. Thus, a detailed discussion of the procedures for relating costs to outcomes is beyond the scope of this chapter. In addition, it is often difficult if not impossible to put a monetary value on the benefits that might result from the implementation and continuance of a program, especially one involving wellness. Some of the basic concepts of a cost analysis can be discovered by reading Chapter 11 of Posavac and Carey's (2010) fine introduction to program evaluation.

A Note on Qualitative Data Analysis

Chapter 3 introduced the difference between a quantitative analysis (numbers involved) and a qualitative analysis (numbers not so critical), and Chapter 10 elaborated upon qualitative analysis of data from non-experimental designs. Although much of the analysis that occurs in program evaluation is quantitative in nature, there is a great deal of qualitative analysis as well, especially in the first three categories of evaluation just described. Thus, during a needs analysis, quantitative data from a community survey and census data can be combined with in-depth interview information from key informants and focus groups. In formative and summative assessments, quantitative data can be supplemented with a qualitative analysis of interviews with agency workers and clients and with direct observations of the program in action. In short, in program evaluation research, it is seldom a question of whether quantitative or qualitative research is better. Although there has been and continues to be debate about the relative merits of quantitative and qualitative evaluation (e.g., Worthen, 2001), thoughtful program evaluators rely on both.

SELF TEST

11.3

1. When are focus groups and community forums used during a program evaluation?
2. What is a formative evaluation and what is the value of one?
3. What is a summative evaluation, and why does it generate more stress than a formative evaluation?

As first mentioned in Chapter 5's discussion of *external validity*, research in psychology is sometimes criticized for avoiding real-world investigations. This chapter on applied research should make it clear that the criticism is without merit. Indeed, concern over application and generalizability of results is not far from the consciousness of all psychologists, even those committed primarily to basic research. It is evident from psychology's history that application is central to American psychology, if for no other reason than Americans can't help it. Looking for practical applications of research is as American as apple pie.

The next chapter introduces a slightly different tradition in psychological research: an emphasis on the intensive study of individuals. As you will see, just as the roots of applied research can be found among psychology's pioneers, experiments with small *N* also trace to the beginnings of the discipline. Before moving on to Chapter 12, however, read Box 11.3, which summarizes some ethical problems likely to be encountered when doing program evaluation research.

BOX 11.3 ETHICS—Evaluation Research and Ethics

Whether evaluating programs that provide services to people, conducting studies in a workplace environment, or evaluating a government service, program evaluation researchers often encounter ethical dilemmas not faced by laboratory psychologists. Some special problems include:

- *Informed consent.* People receiving social services are often powerless. When asked to "volunteer" for a study and sign an informed consent form, they may fear that a failure to sign up could mean a loss of services. In situations like this, researchers must take deliberate steps to reassure participants that no coercion will occur.
- *Maintaining confidentiality.* In some research, confidentiality can be maintained by gathering behavioral data from participants but not adding any personal identifiers. In other studies, however, it is necessary for the researcher to know who the participants are. For instance, the researcher might need to repeatedly contact participants, especially if the study is a longitudinal one, or a researcher might

want to know who replied to a survey so nonrespondents can be contacted again. In such cases, it is important to develop coding systems to protect the identities of participants. Sometimes, participants in longitudinal studies can use aliases, and survey respondents can send back the anonymous survey and a postcard verifying their participation in separate mailings (Sieber, 1998).

- *Perceived injustice.* Some people might object to being in a control group because they could be missing out on some potentially beneficial treatment. Although most control group members in program evaluation research receive the prevailing treatment rather than none at all, control group problems can still happen. For example, *participant crosstalk* (see Chapter 2 and Box 8.2 in Chapter 8) can occur if control group members discover important information about the program being offered to someone else. Their resentment of "special treatment" being given to others can seriously affect the outcome. In a study designed to evaluate worksite changes in a coal mine, for instance, control group miners quickly grew to resent those in the treatment group, whom they felt were getting special attention and did not have to work as hard for the same money (Blumberg & Pringle, 1983). The ill will was even directed at the researchers. Control group workers believed them to be in league with the mine owner in an attempt to break the union. The study as originally designed had to be discontinued.
- *Avoiding conflict with stakeholders.* **Stakeholders** are persons connected with a program in which they have a vested interest, including clients, staff, and program directors. Program evaluators must be aware of and take steps to avoid potential conflict. This means being aware of the needs of stakeholders and explicitly addressing them during all stages of the evaluation.

CHAPTER SUMMARY

Beyond the Laboratory

The goal of applied research is to shed light on the causes of and solutions to real-world problems. Like basic research, however, the outcomes of applied research also contribute to general theories about behavior (e.g., the cognitive interview study contributes to our basic knowledge about the influence of context on memory). American psychologists always have been interested in applied research, partly because of institutional pressures to show the "new" psychological science of the late 19th century could be put to good use. Applied research can encounter ethical problems (e.g., with informed consent) and problems with internal validity (e.g., nonequivalent groups), but it is often strong in external validity.

Quasi-Experimental Designs

Research in which participants cannot be randomly assigned to conditions is referred to as quasi-experimental research. Nonequivalent control group designs are one example. They typically compare pretest/posttest changes in a group receiving some treatment with pre/post changes in a control group formed without random assignment. Regression effects can make interpretation difficult when nonequivalent groups are forced into a degree of equivalency by matching them on pretest scores. In an interrupted time series design, researchers take several measurements both before and after the introduction of the treatment being evaluated. Time series studies enable the researcher to evaluate the effects of trends. Sometimes a nonequivalent control condition, a switching replication, or additional dependent measures can be added to the basic time series design.

Program Evaluation

The field of program evaluation is a branch of applied psychology that provides empirical data about the effectiveness of human service and government programs. Needs analysis studies determine whether a new program should be developed. Census data, surveys, and other community data can help assess need. Formative evaluations determine whether a program is operating according to plan, and summative evaluations assess program outcomes. Cost effectiveness analyses help determine whether a program's benefits are worth the funds invested. Program evaluation research typically combines both quantitative and qualitative methods.

CHAPTER REVIEW QUESTIONS

1. Use the Research Example of traffic signal labeling and food preference and choice as a way of showing how basic research and applied research are related (Trudel et al., 2015).
2. Describe how Hollingworth was able to use fairly sophisticated methodological controls in his applied study of the effects of caffeine.

3. Describe the essential features of a nonequivalent control group design, and explain why Figure 11.2c does not necessarily allow the conclusion that the program was a success.

4. Early program evaluations of Head Start seemed to show that gains made by Head Start children were short-lived; by the third grade, no differences existed between those who had been in the program and those who had not. However, this outcome might have been the result of regression to the mean brought about by the matching procedure used to form the groups. Explain.

5. Describe the Research Example that evaluated whether Play Streets led to increased physical activity in children in terms of why it is a nonexperimental control group, and how researchers tried to equate the groups as much as possible (D'Haese et al., 2015).

6. Describe the essential features of an interrupted time series design and three variations on the basic procedure that can strengthen the conclusions drawn from such a design.

7. Describe two quantitative and two qualitative procedures that can be used when conducting a needs analysis.

8. Distinguish between formative and summative program evaluations. What procedures might be used for each?

9. A finding of "no difference" sometimes occurs in program evaluation research. Explain why this is not necessarily a bad thing.

10. Briefly describe the attributes of the four main types of program evaluation research.

11. Briefly describe the ethical dilemmas that can face people doing program evaluation research.

APPLICATIONS EXERCISES

Exercise 11.1. Identifying Threats to Internal Validity

Threats to internal validity are common in non-experimental studies. What follows is a list of some threats you've encountered in this chapter and in Chapter 5. For each of the hypothetical experiments described, identify which of these threats is most likely to provide a reasonable alternative explanation of the outcome. In some cases, more than one threat could be involved.

Some threats to internal validity:

history	maturation
regression	selection
attrition	selection x history

1. A university dean is upset about the low percentage of freshmen who return to the school as sophomores. Historically, the rate has been around 75%, but in the academic year just begun, only 60% of last year's freshmen return. The dean puts a tutoring program into effect and then claims credit for its effectiveness when the following year's return rate is 65%.

2. Two nearby colleges agree to cooperate in evaluating a new computerized instructional system. College A gets the program and college B doesn't. Midway through the study, college B announces it has filed for bankruptcy (even though it continues to operate). One year later, computer literacy is higher at college A.

3. Twelve women who volunteer for a home birthing program are compared with a random sample of other pregnant women who undergo normal hospital procedures for childbirth. Women in the first group spend an average of 6 hours in labor, while those in the control group spend an average of 9 hours.

4. A 6-week program in managing test anxiety is developed and given to a sample of first-semester college students. Their anxiety levels are significantly lower at the conclusion of the program than they were at the start.

5. A teacher decides to use an innovative teaching technique in which all students will proceed at their own pace throughout the term. The course will have 10 units, and each student goes to unit *N* after completing unit *N* - *1*. Once all 10 units have been completed, the course is over and an A has been earned. Of the initial 30 students enrolled in the class, the final grade distribution looks like this:

16	earned an A
2	failed
12	withdrew from the course during the semester

The instructor considers the new course format an unqualified success.

6. A company decides to introduce a flextime program. It measures productivity for January, runs the program for six months, and then evaluates productivity during the month of June. Productivity increases.

Exercise 11.2. Interpreting Nonequivalent Control Group Studies

A wheel-bearing manufacturer owns two plants, both in Illinois. She wishes to see if money for health costs can be reduced if a wellness program is instituted. One plant (E) is selected for a year-long experimental program that includes health screening and individually tailored fitness activities. The second plant (C) is the nonequivalent control group. Absence-due-to-sickness rates, operationally defined as the number of sick days per year per 100 employees, are measured at the beginning and the end of the experimental year. What follows are four sets of results. Construct a graph for each and decide which (if any) provide evidence of program effectiveness. For those outcomes not supporting the program's effectiveness, provide an alternative explanation for the experimental group's apparent improvement.

Outcome 1	E:	pretest = 125	posttest = 100
	C:	pretest = 125	posttest = 125
Outcome 2	E:	pretest = 125	posttest = 100
	C:	pretest = 100	posttest = 100
Outcome 3	E:	pretest = 125	posttest = 100
	C:	pretest = 130	posttest = 105
Outcome 4	E:	pretest = 125	posttest = 100
	C:	pretest = 110	posttest = 110

Exercise 11.3. Interpreting Time Series Studies

Imagine a time series study evaluating the effects of a helmet law on head injuries among hockey players in amateur city leagues across the nation. Head injuries were significantly lower in the year immediately after the law was passed than in the preceding year. Construct four time series graphs, one for each of the following patterns of results.

1. The helmet law worked.
2. The helmet law seemed to work initially, but its effects were short-lived.
3. The helmet law had no effect; the apparent drop was probably just the result of regression to the mean.
4. The helmet law didn't really work; the apparent drop seemed to reflect a general trend toward reduced violence in the sport.

In the section on interrupted time series designs, we described several variations on the basic design. How might each of those be used to strengthen the hockey study?

Exercise 11.4. Planning a Needs Analysis

You are the head of an advocacy group hired by a school district to develop an anti-bullying program in the public elementary schools in the district. Because you've read this chapter, you respond that a needs analysis should be done. The school superintendent tells you to go ahead and even approves a modest budget for the project. Describe the factors that must be considered before implementing the anti-bullying program in schools and explain the techniques you would use to conduct a needs analysis.

Answers to Self Tests

✓ **11.1**

1. The "dual" functions are solving real-world problems, while contributing to general knowledge about some phenomenon.
2. Miles adapted a basic research methodology, reaction time, to an applied problem, reactions of football linemen.
3. Compared with basic laboratory research, applied research tends to be lower in internal validity and higher in external validity.

✓ **11.2**

1. The groups are nonequivalent; in addition, one group gets one type of treatment, and the other group gets a different treatment (or none at all).
2. Regression.
3. In a switching replication, the treatment program is implemented in two different places and at two different times.

✓ **11.3**

1. During a needs analysis.
2. Formative evaluation assesses a program that is in progress and allows for program improvements to be implemented before the program is completed.
3. Compared to formative evaluations, summative evaluations can eliminate jobs if the result is an ineffective program.