CHAPTER 13

Alternative Research Designs

Although you may think that by now we have covered every conceivable type of research design that psychologists might use to gather data, you would be wrong. There are many other types of research designs. In this chapter we will consider some designs developed by researchers with specific purposes in mind. We will look first at research designs that protect the internal validity of our experiments.

Protecting Internal Validity Revisited

In Chapter 8 we introduced the concept of **internal validity**. The issue of internal validity revolves around **confounding** and **extraneous variables**. When you have an internally valid experiment, you are reasonably certain that your independent variable (IV) is responsible for the changes you observed in your dependent variable (DV). You have established a **cause-and-effect relation**, knowing that the IV *caused* the change in the DV. For example, after many years of painstaking research, medical scientists know that cigarette smoking *causes* lung cancer. Although there are other variables that can trigger cancer, we know that smoking is a causative agent. Our goal as experimenters is to establish similar cause-and-effect relations in psychology. Experiments that are internally valid allow us to make statements such as "X causes Y to occur" with confidence.

Internal validity A type of evaluation of your experiment; it asks whether your IV is the only possible explanation of the results shown for your DV.

Confounding Caused by an uncontrolled extraneous variable that varies systematically with the IV.

Extraneous variables Uncontrolled variables that may unintentionally influence the dependent variable (DV) and thus invalidate an experiment.

Cause-and-effect relation Occurs when we know that a particular IV (cause) leads to specific changes in a DV (effect).

Examining Your Experiment from the Inside

In Chapter 6 we talked about the necessity for controlling extraneous variables in order to reach a clear-cut conclusion from our experiment. It is only when we have designed our experiment in such a way as to avoid the effects of potential extraneous variables that we can feel comfortable about making a cause-and-effect statement; that is, saying that Variable X (our IV) *caused* the change we observed

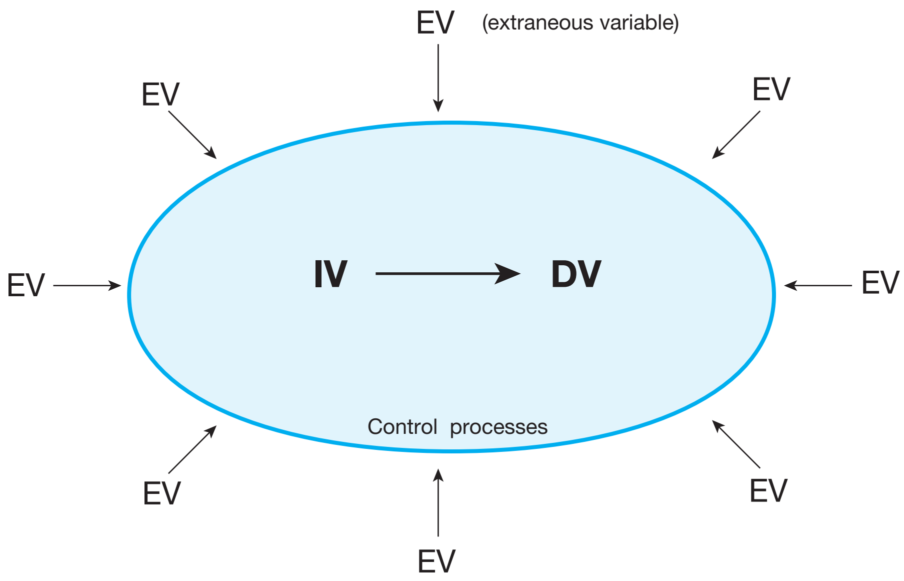


FIGURE 13-1 The Role of Control Processes in Preventing Experimental Confounding.

in Variable Y (our DV). What we are trying to accomplish through our control techniques is to set up a buffer for our IV and DV so that they will not be affected by other variables. This reminds us of a cartoonlike toothpaste commercial we saw—perhaps you have seen it also. When the teeth brushed themselves with the particular brand of toothpaste being advertised, they developed a protective "invisible barrier" against tooth decay. In an analogous manner, our controls give our experiment a barrier against confounding (see Figure 13-1). Similarly, police detectives strive to make their case against a particular suspect airtight. If they have carried out their investigations well, the case against the accused should hold up in court.

Dealing with the internal validity of an experiment is an interesting process. We take many precautions aimed at increasing internal validity as we design and set up our experiment, and we usually evaluate our experiment with regard to internal validity after we have completed the research. If this approach seems a little strange to you, don't be alarmed—it does seem odd at first. Internal validity revolves around the question of whether your IV actually created any change observed in your DV. As you can see in Figure 13-1, if you learned your lessons from Chapter 8 well and used adequate control techniques, your experiment should be free from confounding and you can indeed conclude that your IV caused the change in your DV. Let's review briefly.

PSYCHO-LOGICAL DETECTIVE

Imagine you have been given responsibility for conducting the famous Crest test—you are supposed to determine whether brushing with Crest actually does reduce cavities. Your boss wants you to use an experimental group (Crest) and a control group (Brand X) in the experiment. Write down at least *five* potential extraneous variables for this experiment before reading further.

Were you able to list five possible extraneous variables? The list could be quite long; you may have thought of some possibilities that we didn't. (Although this exercise could easily have been an example dealing with confounding found in Chapter 8, it is also relevant to the issue of internal validity. If you fail to control an important extraneous variable, your experiment will not have internal validity.) Remember, any factor that systematically differs between the two groups (other than the type of toothpaste) could be an extraneous variable that could make it impossible to draw a definite conclusion about the effect of the toothpastes. Here's our (partial) list of possibilities:

number of times brushed per day
amount of time spent in brushing per day
how soon brushing occurs after meals
types of foods eaten
type of toothbrush used
dental genetics inherited from parents
degree of dental care received
different dentists' "operational definition" of what constitutes a cavity
whether the city's water is fluoridated

As we said, this list is not meant to be exhaustive—it merely gives you some ideas of factors that could be extraneous variables. To make certain you understand how an extraneous variable can undermine an experiment's internal validity, let's use an example from the previous list. In addition, we will discover why we take precautions aimed at internal validity *before* the experiment and assess the internal validity of an experiment *afterward*.

When you design the study, you want to make sure that people in the experimental and control groups brush their teeth an equivalent number of times per day. Thus, you would instruct the parents to have their children brush after each meal. Your goal is to have all children brush three times a day. Suppose that you conducted the experiment and gathered your data. When you analyzed the data, you found that the experimental group (Crest) had significantly fewer cavities than the control group (Brand X). Your conclusion seems straightforward at this point: Brushing with Crest reduces cavities compared to brushing with Brand X. As you dig deeper into your data, however, you look at the questionnaire completed by the parents and discover that the children in the experimental group averaged 2.72 brushings a day compared to 1.98 times per day for the children in the control group. Now it is obvious that your two groups differ on two factors: the type of toothpaste used and the number of brushings per day. Which factor is responsible for the lower number of cavities in the experimental group? It is impossible to tell! There is no statistical test that can separate these two confounded factors. You attempted to control the brushing factor before the experiment to assure internal validity, but you could not assess your control technique until after the experiment, when you found out that your experiment was not internally valid. A word to the wise should be sufficient: Good experimental control leads to internally valid experiments.

Remember that we listed nine threats to internal validity in Chapter 8. We also provided you with a variety of control strategies to deal with those threats. Now that you are familiar with research designs, we can explain how some research design strategies eliminate threats to internal validity. As you read about these strategies, you will see that some are a part of those designs discussed in Chapters 10 through 12.

Protecting Internal Validity With Research Designs

There are two approaches you could take to fight the various threats to internal validity. In the first approach you would attempt to come up with nine different answers, one for each threat. Although this approach would be effective in controlling the threats, it would be time consuming and, perhaps, difficult to institute that many different controls simultaneously. Perhaps the idea of controlling the threats through research design occurred to you, even if you could not come up with a specific recommendation. Detectives use standard police procedures to help them protect their cases; experimental design procedures can help us as psychological detectives.

In the three previous chapters we presented you with a variety of experimental designs, often noting various control aspects of those designs; however, we never mentioned the nine general threats to internal validity until this chapter. Can we apply experimental design to these problems? According to Campbell (1957) and Campbell and Stanley (1966), the answer is "yes." Let's take a look at their recommendations.

Random Assignment Although **random assignment** is not a specific experimental design, it is a technique that we can use within our experimental designs. Remember, with random assignment (see Chapter 4) we distribute the experimental participants into our various groups on a random (nonsystematic) basis. Thus, all participants have an equal chance of being assigned to *any* of our treatment groups. The purpose behind random assignment is to create different groups that are equal before beginning our experiment. According to Campbell and Stanley (1966), "[T]he most adequate all-purpose assurance of lack of initial biases between groups is randomization" (p. 25). Thus, random assignment can be a powerful tool. The only drawback to random assignment is that we cannot *guarantee* equality through its use.

Random assignment This control technique ensures that each participant has an equal chance of being assigned to any group in an experiment.

One caution is in order at this point. Because *random* is a frequently used term when dealing with experimental design issues, it sometimes has slightly different meanings. For example, in Chapters 10 through 12 we repeatedly referred to *independent groups* to describe groups of participants that were not correlated in any way (through matching, repeated measures, or natural pairs or sets). It is not unusual to see or hear such independent groups referred to as *random groups*. Although this label makes sense because the groups are unrelated, it is also somewhat misleading. Remember in Chapter 10 when we first talked about matching participants? At that point we stressed that after making your matched pairs of participants, you *randomly assigned* one member of each pair to each group. The same is true of naturally occurring pairs (or sets) of participants. These randomly assigned groups would clearly *not* be independent. Because of the power of random assignment to equate our groups, we should use it at every opportunity. Campbell and Stanley (1966) noted that "within the limits of confidence stated by the tests of significance, randomization can suffice without the pretest" (p. 25). Thus, according to Campbell and Stanley, it may not even be necessary to use matched groups because random assignment can be used to equate the groups.

PSYCHO-LOGICAL DETECTIVE

What is the major exception to Campbell and Stanley's argument that randomization will create equal groups?

We hope that you remembered (from Chapters 10–12) that randomization is supposed to create equal groups *in the long run*. You should be aware of randomization's *possible* shortcoming if you conduct an experiment with small numbers of participants. Although randomization may create equal groups with few participants, we cannot be as confident about this possibility as when we use large groups.

Finally, you should remember from Chapter 4 that random assignment is *not* the same as **random selection**. Random assignment is related to the issue of internal validity; the notion of random selection is more involved with external validity (see Chapter 8).

Experimental Design Campbell and Stanley (1966) reviewed six experimental designs and evaluated them in terms of controlling for internal validity. They recommended three of the designs as being able to control the threats to internal validity we listed in Chapter 8. Let's examine their three recommended designs.

The Pretest–Posttest Control-Group Design The pretest–posttest control-group design appears in Figure 13-2. As you can see, this design consists of two randomly assigned groups of participants, both of which are pretested, with one group receiving the IV.

The threats to internal validity, which we summarized in Chapter 8, are controlled by one of two mechanisms in this design. The random assignment of participants to groups allows us to assume that the two groups are equated before the experiment, thus ruling out **selection** as a problem. Using a pretest and a posttest for *both* groups allows us to control the effects of **history**, **maturation**, and **testing** because they should affect both groups equally. If the control group shows a change between the pretests and posttests, then we know that some factor other than the IV is at work. **Statistical regression** is controlled as long as we assign our experimental and control groups from the same extreme pool of participants. If any of the **interactions with selection** occur, they should affect both groups equally, thus equalizing those effects on internal validity.

The other threats to internal validity are not controlled, but the pretest–posttest control-group design does give us the ability to determine whether they were

R	O_1		O_2	(control group)
R	O_3	X	O_4	(experimental group)

KEY:
R = Random assignment
O = Pretest or posttest observation or measurement
X = Experimental variable or event
Each row represents a different group of participants.
Left-to-right dimension represents passage of time.
Any letters vertical to each other occur simultaneously.
(Note: This key also applies to Figures 13-3, 13-5, and 13-6)

FIGURE 13-2 The Pretest–Posttest Control-Group Design.

Random selection A control technique that ensures that each member of the population has an equal chance of being chosen for an experiment.

Selection A threat to internal validity that can occur if participants are chosen in such a way that the groups are not equal before the experiment; the researcher cannot then be certain that the IV caused any difference observed after the experiment.

History A threat to internal validity; refers to events that occur between the DV measurements in a repeated-measures design.

Maturation An internal validity threat; refers to changes in participants that occur over time during an experiment; could include actual physical maturation or tiredness, boredom, hunger, and so on.

Testing A threat to internal validity that occurs because measuring the DV causes a change in the DV.

Statistical regression This threat to internal validity occurs when low scorers improve or high scorers fall on a second administration of a test solely as a result of statistical reasons.

problematic in a given experiment. We can check to see whether **experimental mortality** was a problem because we measure both groups on two occasions. **Instrumentation** is measured if we are dealing with responses to a test, for example, but it could still remain a problem if human interviewers or observers are used. There is simply no substitute for pretraining when you use humans to record or score data for you. Finally, **diffusion or imitation of treatments** could still remain a problem if participants from the control group (or different experimental groups) learn about the treatments for other groups. Again, though, you do have the control group as a "yardstick" to determine whether their scores increase or decrease in similar fashion to the experimental group's scores. If you see similar changes, you can suspect that internal validity controls may have failed.

Interactions with selection These threats to internal validity can occur if there are systematic differences between or among selected treatment groups based on maturation, history, or instrumentation.

Experimental mortality This threat to internal validity can occur if experimental participants from different groups drop out of the experiment at different rates.

Instrumentation This threat to internal validity occurs if the equipment or human measuring the DV changes its measuring criterion over time.

Diffusion or imitation of treatment This threat to internal validity can occur if participants in one treatment group become familiar with the treatment being received by another group and copy that treatment.

The Solomon Four-Group Design Figure 13-3 contains a diagram of the Solomon four-group design, first proposed by Solomon (1949). Notice that this design is identical to the pretest–posttest control-group design with the first two groups but adds an additional two groups, thus gaining the name *four-group design*. Because the Solomon four-group design has the same two groups as the pretest–posttest control-group design, it has the same protection against the threats to internal validity. The main advantage gained by adding the two additional groups relates to external validity (see Chapter 8).

One problem with the Solomon design comes in conducting statistical analysis of the data because there is no statistical test that can treat all six sets of data at the same time. Campbell and Stanley (1966) suggested treating the posttest scores as a factorial design, as shown in Figure 13-4. Unfortunately, this approach ignores all the pretest scores.

The Posttest-Only Control-Group Design Figure 13-5 shows the posttest-only control-group design. As you can see by comparing Figure 13-5 to Figures 13-2 and 13-3, the posttest-only control-group design is a copy of the pretest–posttest control-group design, without the pretests included, and is a copy of the two added groups in the Solomon four-group design. Does the lack of pretests render the posttest-only control-group design less desirable than the other two designs that include them? No, because we can count on the random assignment to groups to equate the two groups. Thus, using random assignment of participants to groups and withholding the IV from one group to make it a control group is a powerful experimental design that controls the threats to internal validity we covered in Chapter 8.

R	O_1		O_2
R	O_3	X	O_4
R			O_5
R		X	O_6

FIGURE 13-3 The Solomon Four-Group Design. This design is used to protect internal validity.

	No IV	Receives IV
Pretested	O_2	O_4
Unpretested	O_5	O_6

FIGURE 13-4 Factorial Treatment of Solomon Four-Group Design Posttest Scores.

R		O_1
R	X	O_2

FIGURE 13-5 Posttest-Only Control-Group Design. This is a powerful design for protecting internal validity.

PSYCHO-LOGICAL DETECTIVE

After examining Figure 13-5, what type of design (from Chapters 10–12) does this appear to be?

We hope that you identified Figure 13-5 as the two-group design from Chapter 10. We must point out, however, that it is *not* critical to have only two groups in this design. The posttest-only control-group design could be extended by adding additional treatment groups, as shown in Figure 13-6. This extended design should remind you of the multiple-group design discussed in Chapter 11.

Finally, we could create a factorial design from the posttest-only control group by combining two of these designs simultaneously so that we ended up with a block diagram similar to those from Chapter 12.

R		O_1
R	X_1	O_2
R	X_2	O_3
•	•	•
•	•	•
R	X_n	O_{n+1}

FIGURE 13-6 An Extension of the Posttest-Only Control-Group Design. This design permits testing of multiple treatment groups.

It should be clear that the posttest-only design is not defined by the number of groups. What is (are) the defining feature(s) of this design? Take a moment to study Figures 13-5 and 13-6 before answering.

The two features that are necessary to "make" a posttest-only control-group design are random assignment of participants to groups and the inclusion of a control (no-treatment) group. These features allow the design to derive cause-and-effect statements by equating the groups before the experiment and controlling the threats to internal validity.

We hope you can appreciate the amount of control that can be gained by the two simple principles of random assignment and experimental design. Although these principles are simple, they are quite elegant in the power they bring to the experimental situation. You would be wise not to underestimate their importance.

Conclusion

How important is internal validity? It is *the most important* property of any experiment. If you do not concern yourself with the internal validity of your experiment, you are wasting your time. Experiments are intended to produce cause-and-effect statements—to conclude that *X* causes *Y* to occur. If you merely wish to learn something about the association of two variables, you can use one of the nonexperimental methods for acquiring data summarized in Chapter 4 or calculate a correlation coefficient. If you wish to investigate the cause(s) of a phenomenon, you must control any extraneous variables that might affect your dependent variable. You cannot count on your statistical tests to provide the necessary control functions for you. Statistical tests merely analyze the numbers you bring to the test; they do not have the ability to remove confounding effects (or even to discern that confounding has occurred) in your data.

REVIEW SUMMARY

1. One important control for **internal validity** is **random assignment** of participants to groups. This procedure assures us that the groups are equated before beginning the experiment.
2. **Random selection** refers to choosing our participants from a population so that all potential participants could be chosen. Random selection is important to external validity.
3. The pretest–posttest control-group design consists of two groups of participants that have been randomly assigned to an experimental and control group, pretested and posttested, with the experimental group receiving the IV. This design controls for internal validity threats but has the problem of including a pretest.
4. The Solomon four-group design is a copy of the pretest–posttest control-group design except that it adds two groups that have not been pretested. This design also controls for internal validity threats, but there is no statistical test that can be used to analyze all six sets of data.

5. The posttest-only control-group design consists of two groups of participants that are randomly assigned to experimental and control groups, with the experimental group receiving the IV treatment. Both groups are tested with a posttest. This design controls for internal validity threats and is free from other problems.
6. The posttest-only control-group design can be extended to include additional treatment groups or additional IVs.
7. It is essential for an experiment to be internally valid; otherwise, no conclusion can be drawn from the experiment.

Check Your Progress

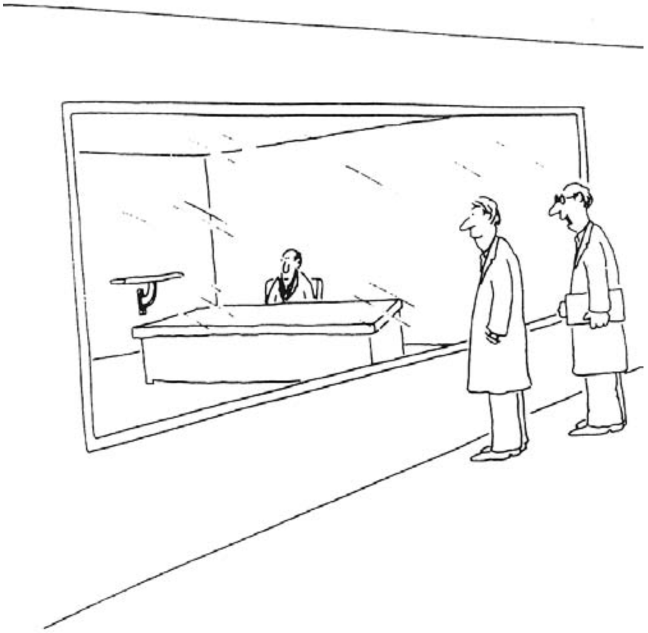
1. The two general methods we use to protect the internal validity of our experiment are ________ and ________.
2. Why is it essential to use random assignment of our participants to their groups?
3. Distinguish between random assignment and random selection.
4. What is the drawback of using the pretest–posttest control-group design to help with internal validity?
5. A friend tells you she was a participant in a psychology experiment and says, "It was crazy! We took a personality test, watched a film, and then took the same test again!" From this description, you could tell that she was in the
   a. control group of a posttest-only control-group design
   b. experimental group of a posttest-only control-group design
   c. control group of a pretest–posttest control-group design
   d. experimental group of a pretest–posttest control-group design
6. What is the drawback of using the Solomon four-group design as a control for internal validity?
7. Diagram the posttest-only control-group design. Why is it a good choice for controlling internal validity?

Single-Case Experimental Designs

A **single-case experimental design** (also known as an $N = 1$ design) is just that. This term simply refers to an experimental design with one participant. This approach, of course, is quite similar to the detective's strategy of pursuing a single suspect.

Single-case experimental design An experiment that consists of one participant (also known as *$N = 1$ designs*).

The $N = 1$ approach probably sounds familiar to you. What data-gathering approach have we studied that involves one participant?



Reprinted with special permission of King Features Syndicate and Tom Cheney.

"Sooner or later he'll learn that when he presses the bar, he'll receive a salary."

Much psychological knowledge has been gained from single-case designs.

We hope you remember the **case-study approach** from Chapter 3. In a case study we conduct an intense observation of a single individual and compile a record of those observations. As we noted in Chapter 3, case studies are often used in clinical settings. If you have taken an abnormal psychology course, you probably remember reading case studies of people with various disorders. The case study is an excellent descriptive technique; if you read a case study about an individual with a mental disorder, you get a vivid picture of that disorder. On the other hand, a case study is *merely* a descriptive or observational approach; the researcher does not manipulate or control variables but simply records observations. Thus, case studies do not allow us to draw cause-and-effect conclusions.

Case-study approach
An observational technique in which a record of observations about a single participant is compiled.

You will remember that we must institute control over the variables in an experiment in order to derive cause-and-effect statements. In a single-case design we institute controls just as we do in a typical experiment—the only difference is that our experiment deals with just one participant. Also, just as in a typical experiment, we must take precautions in dealing with the internal validity of a single-case design. We hope that the single-case design raises many questions for you. After all, it does go against the grain of some of the principles we have developed thus far. Let's take a quick look at this design's history and uses, which will help you understand its importance.

History of Single-Case Experimental Designs

The single-case experimental design has quite an illustrious past in experimental psychology (Hersen, 1982; Hersen & Barlow, 1976). In the 1860s Gustav Fechner explored sensory processes through the use of psychophysical methods. Fechner developed two concepts that you probably remember from your introductory psychology course: *sensory thresholds* and the *just noticeable difference* (*jnd*). Fechner conducted his work on an in-depth basis with a series of individuals. Wilhelm Wundt (founder of the first psychology laboratory) conducted his pioneering work on introspection with highly trained individual participants. Herman Ebbinghaus conducted perhaps the most famous examples of single-case designs in our discipline. Ebbinghaus was the pioneering researcher in the field of verbal learning and memory. His research was unique—not because he used the single-case design, but because he was the single participant in those designs. According to Dukes (1965), Ebbinghaus learned about 2,000 lists of nonsense syllables in his research over many years. Dukes provided several other examples of famous single-case designs with which you are probably familiar, such as Cannon's study of stomach contractions and hunger, Watson and Rayner's study of Little Albert's learned fears, and several researchers' work with language learning in individual apes.

Other than the ape-language studies cited by Dukes (1965), all these single-case design examples date to the 1800s and early 1900s. Dukes found only 246 single-case examples in the literature between 1939 and 1963. Clearly, there are fewer examples of single-case designs than group designs in the literature.

PSYCHOLOGICAL DETECTIVE

Can you think of a reason why single-case designs may have been more popular in the past?

Hersen (1982) attributed the preference for group designs over single-case designs to statistical innovations made by Sir Ronald A. Fisher. Fisher was a pioneer of many statistical approaches and techniques. Most important for this discussion, in the 1920s he developed analysis of variance (ANOVA; Spatz, 2001), which we covered in detail in Chapters 11 and 12. Combined with Gosset's early 1900s development of a test based on the *t* distribution (see Chapter 10), Fisher's work gave researchers a set of inferential statistical methods with which to analyze sets of data and draw conclusions. You may have taken these tests for granted and assumed that they had been around forever, but that is not the case. As these methods became popular and accessible to more researchers, the use of single-case designs declined. In today's research world, statistical analyses of incredibly complex designs can be completed in minutes (or even seconds) on computers you can hold in your hand. The ease of these calculations has probably contributed to the popularity of group designs over single-case designs.

Experimental analysis of behavior A research approach popularized by B. F. Skinner, in which a single participant is studied.

Uses of Single-Case Experimental Designs

There are still some researchers who use single-case designs. Founded by B. F. Skinner, the **experimental analysis of behavior** approach continues to employ this technique. Skinner (1966) summarized his philosophy in this manner: "Instead of studying a thousand rats for one hour each, or a hundred rats for ten hours each, the investigator is likely to study one rat for a thousand hours" (p. 21). The Society for the Experimental Analysis of Behavior was formed and began publishing its own journals, the *Journal of the Experimental Analysis of Behavior* (in 1958) and the *Journal of Applied Behavior Analysis* (in 1968). Single-case designs are thus still used today; however, the number of users is small compared to those who use group designs, as you could guess by the handful of journal titles devoted to this approach.

One question that might occur to you is "Why use a single-case design in the first place?" Sherlock Holmes knew that "the world is full of obvious things which nobody by any chance ever observes" (Doyle, 1927, p. 745). Dukes (1965) provided a number of convincing arguments for and situations that require single-case designs. Let's look at several. First, a sample of one is all you can manage if that sample exhausts the population. If you have access to a participant who is unique, you simply cannot find other participants. Of course, this example is perhaps closer to a case study than to an experiment because there would be no larger population to which you could generalize your findings. Second, if you can assume perfect generalizability, then a sample of one is appropriate. If there is only inconsequential variability among members of the population on a particular variable, then measuring one participant should be sufficient. Third, a single-case design would be most appropriate when a single *negative* instance would refute a theory or an assumed universal relation. If the scientific community believes that "reinforcement always increases responding," then finding one instance in which reinforcement does *not* increase responding invalidates the thesis. Fourth, you may simply have limitations on your opportunity to observe a particular behavior. Behaviors in the real world (i.e., nonlaboratory behaviors) may be so rare that you can locate only one participant who exhibits the behavior. Dukes used examples of people who feel no pain, who are totally color-blind, or who exhibit dissociative identity disorder (again, close to a case study). You may remember reading about H. M. when you studied memory in introductory psychology. Because of the surgery for epilepsy that removed part of his brain, H. M. could no longer form new long-term memories. Researchers have studied H. M. for almost 50 years for clues about how the brain forms new memories (Corkin, 1984; Hilts, 1995). H.M.'s case was famous enough that the *New York Times* carried his obituary when he died in late 2008 (http://www.nytimes.com/2008/12/05/us/05hm.html?_r=2). Fifth, when research is extremely time consuming and expensive, requires extensive training, or presents difficulties with control, an investigator may choose to study just one participant. The studies in which researchers have attempted to teach apes to communicate through sign language, plastic symbols, or computers fall into this category. Obviously, there are instances in which a single-case design is totally appropriate.

General Procedures of Single-Case Experimental Designs

Hersen (1982) listed three procedures that are characteristic of single-case designs: repeated measures, baseline measurement, and changing one variable at a time. Let's see why each of these procedures is important.

Repeated Measures When we deal with many participants, we often measure them only once and then average all our observations. When you are dealing with only one participant, however, it is important to make sure that the behavior you are measuring is consistent. You would therefore repeatedly measure the participant's behavior. Control during the measurement process is extremely important. Hersen and Barlow (1976) noted that the procedures for measurement "must be clearly specified, observable, public, and replicable in all respects" (p. 71). In addition, these repeated measurements "must be done under exacting and totally standardized conditions with respect to measurement devices used, personnel involved, time or times of day measurements are recorded, instructions to the subject, and the specific environmental conditions" (p. 71). Thus, conducting a single-case experiment and making repeated measurements do *not* remove the experimenter's need to control factors as carefully as possible.

Baseline A measurement of a behavior made under normal conditions (i.e., no IV is present); a control condition.

Baseline Measurement In most single-case designs the initial experimental period is devoted to determining the **baseline** level of behavior. In essence, baseline measurement serves as the control condition against which to compare the behavior as affected by the IV. When you are collecting baseline data, you hope to find a stable pattern of behavior so that you can more easily observe any change that occurs in the behavior after your intervention (IV). Barlow and Hersen (1973) recommended that you collect *at least* three observations during the baseline period in order to establish a trend in the data. Although you may not achieve a stable measurement, the more observations you have, the more confident you can be that you have determined the general trend of the observations. Figure 13-7 depicts a hypothetical stable baseline presented by Hersen and Barlow (1976). Notice that they increased their odds of finding a stable pattern by collecting data three times per day and averaging those data for the daily entry.

Changing One Variable at a Time In a single-case design it is vital that, as the experimenter, you change only one variable at a time when you move from one phase of the experiment to the next.

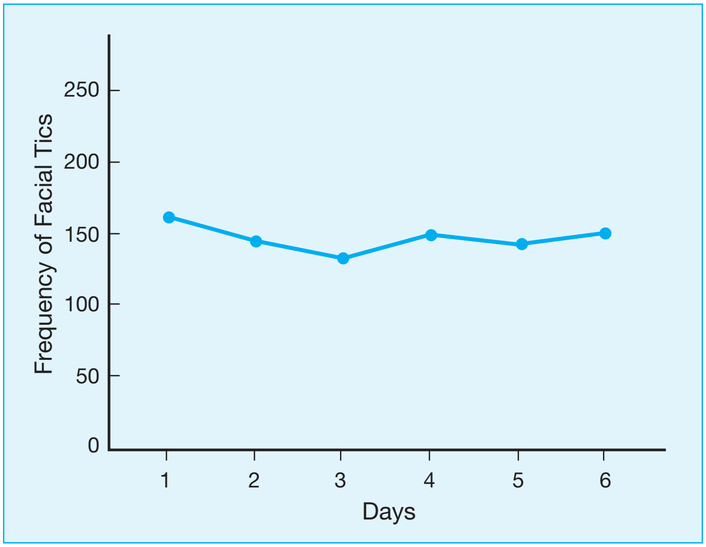


FIGURE 13-7 A Stable Baseline. Hypothetical data for mean number of facial tics averaged over three daily 15-minute videotaped sessions.

Source: Figure 3-1 from *Single-Case Experimental Designs: Strategies for Studying Behavioral Change,* by M. Hersen and D. H. Barlow, 1976, New York: Pergamon Press, p. 77. Used with permission of the publisher.

Why would it be important to change only one variable at a time in a single-case design?

We hope that the answer to this question came easily. Changing one variable at a time is a basic experimental control procedure that we have stressed many times. If you allow two variables to change simultaneously, then you have a confounded experiment and cannot tell which variable has caused the change in the behavior that you observe. This situation is exactly the same in a single-case design. If you record your baseline measurement, change several aspects of the participant's environment, and then observe the behavior again, you have no way of knowing which changed aspect affected the behavior.

Statistics and Single-Case Experimental Designs

Traditionally, researchers have not computed statistical analyses of results from single-case designs. Not only has the development of statistical tests for such designs lagged behind multiple-case analyses, but also there is controversy about *whether* statistical analyses of single-case designs are even appropriate (Kazdin, 1976). Both Kazdin (1976) and Hersen (1982) summarized the arguments concerning statistical analyses. Let's take a quick look at this controversy.

The Case Against Statistical Analysis As we mentioned, tradition and history say that statistical analyses are not necessary in single-case designs. The tradition has been to inspect visually ("eyeball") the data to determine whether change has taken place. Researchers who hold this position believe that treatments that do not produce visually apparent effects are either weak or ineffective. Skinner (1966) wrote that "rate of responding and changes in rate can be directly observed . . . [and] statistical methods are unnecessary" (p. 20).

Because many single-case studies involve clinical treatments, another argument against statistical analysis is that statistical significance is not always the same as clinical significance. A statistical demonstration of change may not be satisfying for practical application. "For example, an autistic child may hit himself in the head 100 times an hour. Treatment may reduce this to 50 times per hour. Even though change has been achieved, a much larger change is needed to eliminate behavior" (Kazdin, 1984, p. 89).

Finally, to the pro-statistics folks who argue that statistical analyses may help find effects that visual inspection would not (see next section), the anti-statistics camp makes the point that such subtle effects may not be replicable (Kazdin, 1976). As you saw in Chapter 8, if you cannot replicate a result, it has no external validity.

The Case for Statistical Analysis The argument for using statistical analyses of single-case designs revolves primarily around increased accuracy of conclusions. Jones, Vaught, and Weinrott (1977) have provided the most persuasive appeal for such analyses. They reviewed a number of studies published in the *Journal of Applied Behavior Analysis* that used visual inspection of data to draw conclusions. Jones et al. found that analyses of these data showed that sometimes conclusions drawn from visual inspections were correct and that sometimes the conclusions were incorrect. In the latter category both Type I and Type II errors (see Chapter 9) occurred. In other words, some statistical analyses showed *no* effect when the researchers had said there was an effect, and some analyses showed significant effects when the researchers had said there were none. Kazdin (1976) pointed out that statistical analyses are particularly likely to uncover findings that do not show up in visual inspection when a stable baseline is not established, new areas of research are being investigated, or testing is done in the real world, which tends to increase extraneous variation.

As you can tell, there is no clear-cut answer concerning the use of statistics with single-case designs. Most researchers probably make their decision in such a situation based on a combination of personal preference, the audience for the information, and potential journal editors. Covering the various tests used to analyze single-case designs is beyond the scope of this text. Adaptations of *t* tests and ANOVA have been used, but these approaches have suffered from some problems. For further information about such tests, see Kazdin (1976).

Representative Single-Case Experimental Designs

Researchers use a standard notation for single-case designs that makes the information easier to present and conceptualize. In this notation, **A** refers to the baseline measurement and **B** refers to the measurement during or after treatment. We read the notation for single-case designs from left to right, to denote the passage of time.

A Refers to the baseline measurement in a single-case design.

B Refers to the outcome (treatment) measurement in a single-case design.

A-B Design In the **A-B design**, the simplest of the single-case designs, we make baseline measurements, apply a treatment, and then take a second set of measurements. We compare the B (treatment) measurements to the A (baseline) measurements in order to determine whether a change has occurred. This design should remind you of a pretest–posttest design except for the absence of a control group. In the A-B design, the participant's A measurements serve as the control for the B measurements.

A-B design A single-case design in which you measure the baseline behavior, institute a treatment, and use a posttest.

For example, Hall et al. (1971) used this approach in a special-education setting. A 10-year-old boy (Johnny) continually talked out and disrupted the class, which led other children to imitate him. The researchers asked the teacher to measure Johnny's baseline talking-out behavior (A) for five 15-minute sessions under normal conditions. In implementing the treatment (B), the teacher ignored the talking out and paid more attention to Johnny's productive behavior (attention was *contingent* on the desired behavior), again for five 15-minute sessions. Johnny's talking out diminished noticeably.

Hersen (1982) rated the A-B design as one of the weakest for inferring causality and noted that it is often deemed correlational.

Why do you think the A-B design is weak concerning causality?

The A-B design is poor for determining causality because of many of the threats to internal validity that we saw in Chapter 8. It is possible that another factor could vary along with the treatment. This possibility is especially strong for any extraneous variables that could be linked to time passage, such as history, maturation, and instrumentation. If such a factor varied across time with the treatment, then any change in B could be due to *either* the treatment or the extraneous factor. Because there is no control group, we cannot rule out the extraneous variable as a causative factor.

Can you think of a solution to the causality problem inherent in the A-B design? Remember that you cannot add a control group *or* participants because this is a single-case design. Any control must occur with the single participant.

The solution to this causality problem requires us to examine our next single-case design.

A-B-A Design In the **A-B-A design**, the treatment phase is followed by a return to the baseline condition. If a change in behavior during B is actually due to the experimental treatment, the change should disappear when B is removed and you return to the baseline condition. If, on the other hand, a change in B was due to some extraneous variable, the change will not disappear when B is removed. Thus, the A-B-A design allows a causal relation to be drawn.

In Hall et al.'s (1971) experiment, the teacher did return to the baseline condition with Johnny. When the teacher began again to pay attention to Johnny's talking-out behavior, that behavior increased considerably. This return to the previous behavior strengthened the researchers' claim that the treatment had caused the original decrease in Johnny's talking out.

A-B-A design A single-case design consisting of a baseline measurement, a treatment, a posttest, and a return to the baseline condition. It may not be recommended if the participant is left without a beneficial or necessary treatment in the second baseline.

There is one glaring drawback to the A-B-A design. Think about the implications of conducting a baseline–treatment–baseline experiment. Can you spot the drawback? How would you remedy this problem?

If you end your experiment on an A phase, this leaves the participant in a baseline condition. If the treatment is a beneficial one, the participant is "left hanging" without the treatment. The solution to this problem requires us to examine another single-case design.

On the other hand, returning to an A phase can give the researcher an idea of how effective a treatment was. Aurelie Welterlin (2004), a student at the University of North Carolina Chapel Hill, worked with a 7-year-old boy diagnosed with autism who exhibited impaired social interaction skills. Welterlin had the boy and two female peers play in a room together and measured the number of times the boy interacted with the peers. During the baseline

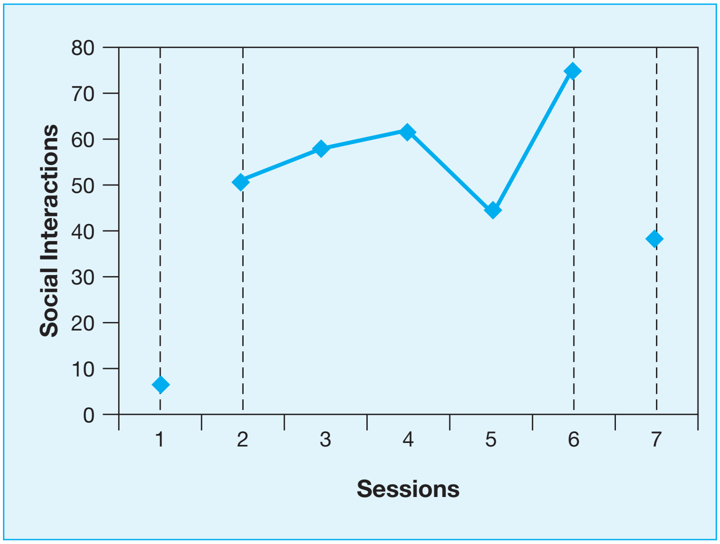


FIGURE 13-8 A-B-A Design. Teaching a 7-Year-Old Boy to Increase Social Interaction. Sessions 1 and 7 represent baseline periods; Sessions 2–6 used cuing to prompt interaction.

Source: Adapted from "Social Initiation in Children with Autism: A Peer Playgroup Intervention," by A. Welterlin, 2004, *Psi Chi Journal of Undergraduate Research*, *9,* pp. 97–104. Copyright © 2004 Psi Chi, The National Honor Society in Psychology (www.psichi.org). Reprinted by permission. All rights reserved.

period (A; Session 1), it was highly unusual for the boy to interact at all (see Figure 13-8). During the intervention sessions (B; Sessions 2–6), a facilitator cued the boy in an attempt to get him to interact with the girls. During the second baseline (A; Session 7), the facilitator did not provide any cues to the boy. An examination of Figure 13-8 shows that the intervention (cuing) did produce more social interaction from the boy. The second baseline period shows that, without the cues, the boy's social interaction decreased; however, it remained higher than it had been in the original baseline. Welterlin was thus able to demonstrate that cuing did increase social interaction (B) and that its effects persisted even in the absence of the cuing (second baseline).

A-B-A-B design A single-case design consisting of a baseline, treatment, posttest, return to baseline, repeated treatment, and second posttest. This design gives the best chance of isolating causation.

A-B-A-B Design As you can figure out by now, the **A-B-A-B design** begins with a baseline period followed by treatment, baseline, and treatment periods consecutively. This design adds a final treatment period to the A-B-A design, thereby completing the experimental cycle with the participant in a treatment phase. Hersen and Barlow (1976) pointed out that this design gives two transitions (B to A and A to B), which can demonstrate the effect of the treatment variable. Thus, our ability to draw a cause-and-effect conclusion is further strengthened.

Hall et al. (1971) actually used the A-B-A-B design in their experiment with Johnny. After measuring Johnny's baseline talking-out behavior (A) under normal conditions, the teacher implemented the treatment (B) by ignoring the talking out and paying attention only to Johnny's productive behavior. The teacher then repeated the A and B phases. Results from this study appear in Figure 13-9. This graph shows us several things. First, visual inspection of these results should be enough to convince us of the efficacy of the treatment—the difference between baseline and treatment conditions is dramatic. This graph is a good

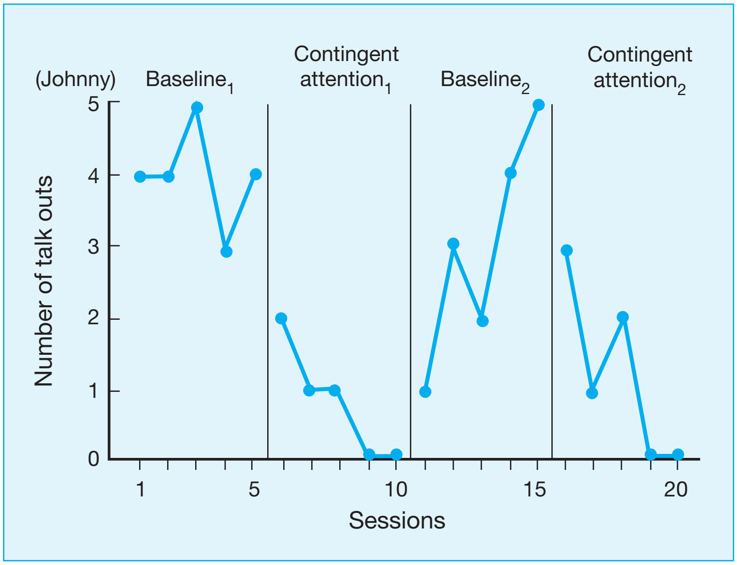


Figure 13-9 Talking-Out Behavior in a Mentally Retarded Student. A record of talking-out behavior of an educable mentally retarded student. **Baseline$_1$**: before experimental conditions; **Contingent Attention$_1$**: systematic ignoring of talking out and increased teacher attention to appropriate behavior; **Baseline$_2$**: reinstatement of teacher attention to talking-out behavior; **Contingent Attention$_2$**: return to systematic ignoring of talking out and increased attention to appropriate behavior.

Source: Figure 2 from "The Teacher as Observer and Experimenter in the Modification of Disrupting and Talking-out Behaviors," by R. V. Hall, R. Fox, D. Willard, L. Goldsmith, M. Emerson, M. Owen, et al., 1971, *Journal of Applied Behavior Analysis*, *4*, p. 143.

illustration of why many researchers who use single-case designs believe that statistics are unnecessary. Second, it is apparent that the treatment did work. When the teacher stopped attending to Johnny's talking-out behavior and paid attention to his productive behavior, the talking out decreased substantially. Third, we can determine that the increased productive behavior was caused by the contingent attention because of the rapid increase in talking out when the attention was removed (see Baseline$_2$ in Figure 13-9).

Design and the Real World From the preceding sections it should be clear that the A-B-A-B design is the preferred design for single-case research; however, we must ask whether typical practice actually follows the recommended path. Hersen and Barlow (1976) acknowledged that researchers often use the A-B design despite its shortcomings in terms of demonstrating causality. The main reason the A-B design is used concerns either the inability or undesirability to return to the baseline in the third stage. In the real world, perfect experimental design cannot always be used. We must simply accept that our ability to draw definitive conclusions in such instances is limited. Let's look at three common situations that preclude using a design other than the A-B design.

First, as is typical in many field experiments, it may be impractical to reverse a treatment. Campbell (1969, p. 410) urged politicians to conduct social reforms as experiments, proposing that they initiate a new policy on an experimental basis. If after five years there had been no significant improvement, he recommended that the politicians shift to a different policy.

Political realities, of course, would not allow social change to be conducted experimentally. Campbell provided a good example of this problem. In 1955 Connecticut experienced a record number of traffic fatalities. The governor instituted a speeding crackdown in 1956, and traffic fatalities fell by more than 12%. Once this result occurred, it would have been politically stupid for the governor to announce, "We wish to determine whether the speeding crackdown actually caused the drop in auto deaths. Therefore, in 1957 we will relax our enforcement of speeding laws to find out whether fatalities increase once again." Yet this change is what would be necessary in order to rule out rival hypotheses and draw a definitive cause-and-effect statement.

Second, it may be unethical to reverse a treatment. Lang and Melamed (1969) worked with a 9-month-old boy (see Figure 13-10) who had begun vomiting after meals when he was about 6 months old. Doctors had implemented dietary changes, conducted medical tests, performed exploratory surgery, but could find no organic cause. The boy weighed 9 pounds, 4 ounces at birth, grew to 17 pounds at 6 months of age, but weighed only 12 pounds at 9 months. The child was being fed through a nose tube and was in critical condition (see Figure 13-10A). Lang and Melamed instituted a treatment consisting of brief and repeated shocks applied to the boy's leg at the first signs of vomiting and ending when vomiting ceased. By the third treatment session, one or two brief shocks were enough to stop the vomiting. By the fourth day of

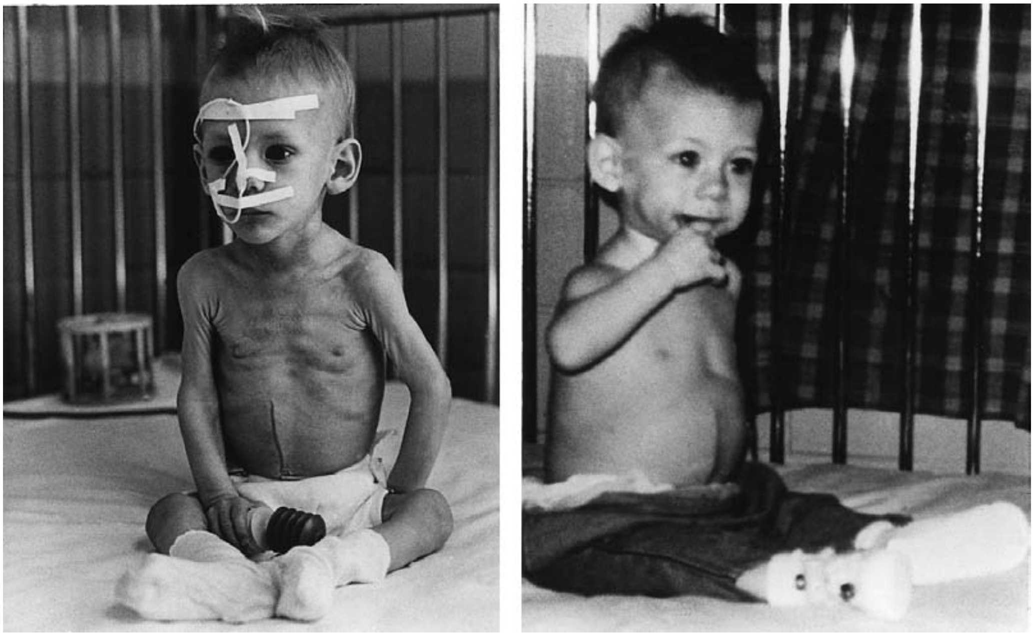


FIGURE 13-10 Nine-Month-Old Boy Hospitalized for Frequent Vomiting (**A**) Before Treatment and (**B**) After Treatment (13 Days Later). The photograph at the left was taken during the observation period just prior to treatment. (It clearly illustrates the patient's debilitated condition: lack of body fat, skin hanging in loose folds. The tape around the face holds tubing for the nasogastric pump. The photograph at the right was taken on the day of discharge from the hospital, 13 days after the first photo. The 26% increase in body weight already attained is easily seen in the full, more infantlike face, the rounded arms, and more substantial trunk.)

Source: Figure 1 from "Avoidance Conditioning Therapy of an Infant with Chronic Ruminative Vomiting," by P. J. Lang and B. G. Melamed, 1969, *Journal of Abnormal Psychology*, 74, pp. 1–8.

treatment, vomiting stopped and treatment was discontinued. Two days later, some vomiting occurred, so the procedure was reinstated for three sessions. Five days later, the child was dismissed from the hospital (see Figure 13-10B). A month later, he weighed 21 pounds, and 5 months later weighed over 26 pounds, with no recurrence of vomiting. Although this treatment bears some resemblance to an A-B-A-B design (because of the brief relapse), the additional session was not originally intended and was *not* conducted as an intentional removal of B to chart a new baseline—the researchers believed that the problem had been cured at the point treatment was discontinued. We are certain that you can see why ethical considerations would dictate an A-B design in this instance rather than the more experimentally rigorous A-B-A-B design.

Finally, it may be impossible, undesirable, or unethical to reverse a treatment if learning takes place during the treatment. Bobby Traffanstedt (1998), a student at the University of Central Arkansas in Conway, used an A-B design to modify a 10-year-old boy's TV watching and exercise behaviors. Traffanstedt wanted to teach the boy to spend less time watching TV and more time exercising. He used the operant procedures of shaping and reinforcement while working with the child for several weeks. The baseline (Week 1) and posttest (Weeks 2–9) behavior measures appear in Figure 13-11. As you can see, visual inspection of these data is convincing.

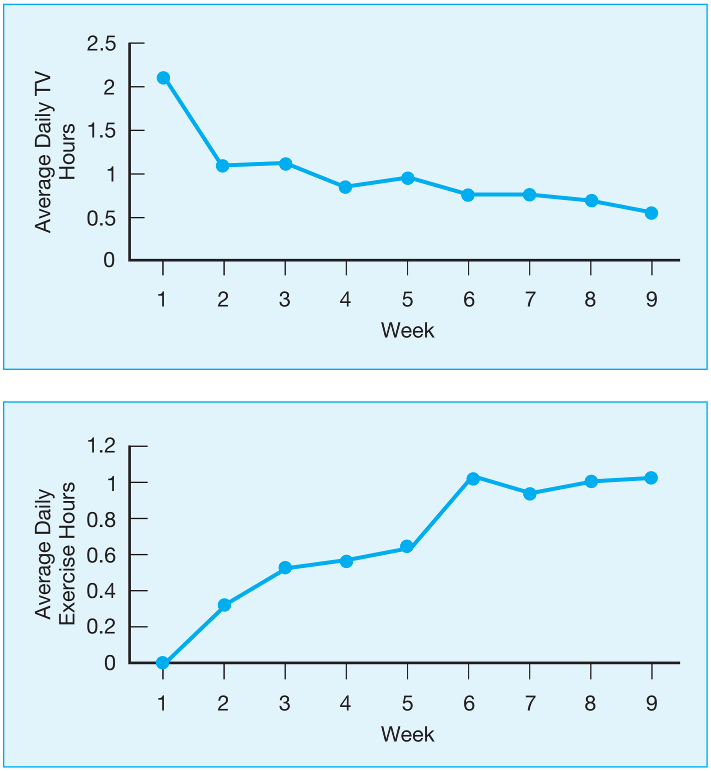


Figure 13-11 Teaching a 10-Year-Old Boy to Decrease TV Viewing and Increase Exercise. Traffanstedt (1998) used shaping and reinforcement to modify the boy's behavior.

Source: Adapted from "Weight Reduction Using Behavior Modification" by B. Traffenstedt, 1998, *Journal of Psychological Inquiry*, 3, pp. 19–23.

PSYCHOLOGICAL DETECTIVE

Why did Traffanstedt (1988) *not* use the A-B-A-B design in this experiment?

Because Traffanstedt had successfully taught the child to spend less time watching TV and more time exercising, he did not want to "undo" this learning and return to the baseline condition. Traffanstedt had already attempted to fade out the reinforcement gradually over the course of Weeks 2 through 9; going back to baseline was not really feasible. Having learned the new behaviors, it would make no sense to return the boy to the baseline condition.

The conclusion to this section is that you as an experimenter may find yourself caught in the middle. On the one hand, you have the knowledge of proper experimental design and what is necessary to yield cause-and-effect explanations. On the other hand, you have the realities of applied situations. The best rule of thumb for such situations is that you should use the most stringent experimental design you can, but you should not give up on an important project if you cannot use the absolute best design that exists. As a psychological detective, you have an edge on the real-life detective, who cannot apply a design even as rigorous as those we have presented in this section. The police detective must always work on a solution after the fact.

Additional Single-Case Designs In presenting the A-B, A-B-A, and A-B-A-B designs, we have merely scratched the surface of single-case designs. We have covered the designs we think you might be likely to use in the near future. As our references show, entire books have been written about single-case designs. Hersen and Barlow (1976) covered many additional variations on single-case designs, including designs with multiple baselines, multiple schedules, and interactions. Thus, if you ever envision a single-case design that is more complicated than the ones we have presented in this text, we refer you to Hersen and Barlow or a similar book dealing with single-case designs.

REVIEW SUMMARY

1. **Single-case experimental designs** are experiments that deal with a single participant.
2. Single-case designs have several legitimate uses.
3. Single-case designs are characterized by repeated measures, **baseline** measurement, and changing one variable at a time.
4. There is controversy over the use of statistics with single-case designs. The traditional approach has been to draw conclusions by visually examining the data. Proponents of statistical analysis maintain that analysis yields more accurate conclusions.
5. The **A-B-A-B** single-case design allows you the best chance to draw a cause-and-effect conclusion regarding a treatment. Realities of the real world often force the use of **A-B** designs, which are particularly prone to alternative explanations.

Check Your Progress

1. Why were single-case designs quite popular in psychology's early years but less popular today?
2. How can a single-case design be used to disprove a theory?
3. To come up with a comparison in the single-case design, we first measure behavior before the treatment during the ________ period. To get a stable measurement, we should make at least ________ observations.
4. In essence, ________ serve(s) as the control condition in the single-case design.
   a. baseline measurements
   b. repeated measures
   c. changing one variable at a time
   d. experimental analysis of behavior
5. Summarize two arguments for and two arguments against the use of statistical analysis in single-case designs.
6. Match the design with the appropriate characteristic.

1. A-B	A. leaves the participant in a baseline phase
2. A-B-A	B. best single-case design for determining cause-and-effect relations
3. A-B-A-B	C. has many threats to internal validity

7. Why might you be forced to use an A-B single-case design in the real world? Give an original example of such a situation.

Quasi-Experimental Designs

In this section we will deal with designs that are virtually identical to true experimental designs with the *exception* of random assignment of participants to groups. When we are able to manipulate an IV and measure a DV but *cannot* randomly assign our participants to groups, we must use a **quasi-experimental design**. Similarly, police detectives sometimes face situations in which they must build their case on circumstantial evidence rather than on direct evidence.

Quasi-experimental design A research design used when the researcher cannot randomly assign experimental participants to the groups but the researcher does manipulate an IV and measure a DV.

PSYCHO-LOGICAL DETECTIVE

What problem results when we cannot randomly assign research participants to groups?

Not being able to assign our participants randomly to their groups has the effect of violating an important assumption that allows us to draw cause-and-effect conclusions from our experiments—the assumption of equal groups before the experiment. Even if we can randomly select participants from a larger group, we cannot make cause-and-effect statements without random assignment. For example, you could randomly *select* students from an introductory psychology course, but you could not randomly *assign* them to groups based on sex! As Campbell and Stanley (1966) pointed out, the assumption of random assignment has been an important part of statistics and experimental design since the time of Fisher. If we unknowingly began an experiment with unequal groups and our statistics showed a difference after the experiment, we would make a Type I error (see Chapter 9) by concluding that the IV caused the difference that was actually present from the outset. Clearly, this conclusion could be wrong.

It is likely that our description of quasi-experimental design reminds you of the ex post facto studies we covered in Chapter 4. Some writers categorize ex post facto and quasi-experimental designs together and some separate them. We will draw a small, but significant, distinction between the two. Remember, in Chapter 4 we described the ex post facto study as having an IV that had *already* occurred and could not be manipulated. Thus, if we wish to study sex differences on mathematics or English achievement, we are studying the IV of biological sex, which we cannot control or manipulate. Of course, because the IV is a preexisting condition, we also cannot randomly assign our participants to groups.

On the other hand, in a quasi-experimental design our participants belong to preexisting groups that cannot be randomly assigned; however, we *do* have control over the IV—we can administer it when and to whom we wish. Thus, we could choose our participants on the basis of sex and *then* have some of them participate in a workshop designed to improve their math or English achievement. In this case the workshop (or lack thereof) would serve as the IV for the preexisting groups of boys and girls, and the math or English achievement scores would be the DV. Obviously, random assignment is impossible in this case. Quasi-experimental designs are a step closer to true experimental designs than ex post facto studies because you, as the experimenter, are able to exert control over the IV and its administration. Being able to administer your own IV is preferable to having nature administer it for you, at least in terms of control.

The basic rationale for using quasi-experimental designs is the same as that for ex post facto studies—your inability to assign participants at random. According to Hedrick, Bickman, and Rog (1993), "[A] quasi-experimental design is not the method of choice, but rather a fall-back strategy for situations in which random assignment is not possible" (p. 62). When dealing with selection variables that do not allow for random assignment, we have the choice of using a quasi-experimental design or simply ignoring an important or interesting experimental question. Instead of letting such questions go unasked, researchers resort to quasi-experimental research.

History of Quasi-Experimental Designs

It is difficult to trace the history of quasi-experimental designs. Although McGuigan (1960) did not include the term in the first edition of his classic experimental psychology text, Campbell and Stanley did use it in the title of their 1966 guide to experimental design. There is little doubt, however, that researchers were tackling quasi-experimental design problems long before Campbell and Stanley's published work. Cook and Campbell (1979) noted that some researchers were writing about quasi-experiments in the 1950s, although the term did not originate until later. It is likely that Campbell and Stanley (1966) and Cook and Campbell (1979) are responsible for elevating quasi-experimental work to the respectable position it holds today.

Uses of Quasi-Experimental Designs

Hedrick et al. (1993) listed several specific situations that require quasi-experimental designs. Let's take a brief look at their list. First, there are many variables that simply make random assignment impossible. If we wish to study participants from certain groups (e.g., based on sex, age, previous life experiences, personality characteristics), we must use quasi-experimental designs. Second, when you wish to evaluate an ongoing program or intervention (a retrospective study), you would have to use a quasi-experimental design. Because the program began before you decided to evaluate it, you would have been unable to use control procedures from the outset. Third, studies of social conditions demand quasi-experimental designs. You would not study the effects of poverty, race, unemployment, or other such social factors through random assignment. Fourth, it is sometimes the case that random assignment is not possible because of expense, time, or monitoring difficulties. For example, if you conducted a cross-cultural research project involving participants from several different countries, it would be nearly impossible to guarantee that the same random assignment procedures were used in each setting. Fifth, the ethics of an experimental situation, particularly with psychological research, may necessitate quasi-experimentation. For example, if you are conducting a research program to evaluate a certain treatment, you must worry about the ethics of withholding that treatment from people who could benefit from it. As you will see, quasi-experimentation provides a design that will work in such situations to remove this ethical dilemma.

Representative Quasi-Experimental Designs

Unlike the single-case design, we do not include sections covering general procedures and statistics of quasi-experimental designs. It is difficult to derive general principles because the representative designs we are about to introduce are so varied in nature. Because quasi-experimental designs resemble true experiments, the use of statistics for quasi-experimental designs is not an issue; the traditional statistical tests used with true experiments are also appropriate for quasi-experiments.

Nonequivalent Group Design The **nonequivalent group design** (Campbell & Stanley, 1966) appears in Figure 13-12.

Nonequivalent group design A design involving two or more groups that are not randomly assigned; a comparison group (no treatment) is compared to one or more treatment groups.

O_1		O_2	(comparison group)
O_1	X	O_2	(treatment group)

KEY:
R = Random assignment
O = Pretest or posttest observation or measurement
X = Experimental variable or event
Each row represents a different group of participants.
Left-to-right dimension represents passage of time.
Any letters vertical to each other occur simultaneously.
Note: This key also applies to Figures 13-15, and 13-18.

FIGURE 13-12 The Nonequivalent Group Design.

PSYCHOLOGICAL DETECTIVE

The nonequivalent group design should remind you of a design that we covered in the section on research designs that protect internal validity. Which design does it resemble? How is it different? What is the implication of this difference?

If you turn back to Figure 13-2, you will see that the nonequivalent group design bears a distinct resemblance to the pretest–posttest control-group design; however, the nonequivalent group design is missing the *R*s in front of the two groups; random assignment is *not* used in creating the groups. The lack of random assignment means that our groups may differ before the experiment—thus the name *nonequivalent group* design.

You also will notice that the two groups are labeled as the *comparison group* (rather than control group) and the *treatment group* (rather than experimental group [from Hedrick et al., 1993]). The *treatment* to *experimental* change is not particularly important; those terms could be used interchangeably; however, changing the name from *control* to *comparison* group is important and meaningful. In the nonequivalent group design, this group serves as the comparison to the treatment group but cannot truly be called a control group because of the lack of random assignment.

It is possible to extend the nonequivalent group design to include more than one treatment group if you wish to contrast two or more treatment groups with your comparison group. The key to the nonequivalent group design is creating a good comparison group. As far as is possible, we attempt to create an equal group through our selection criteria rather than through random assignment.

> Examples of procedures for creating such a group include using members of a waiting list for a program/service; using people who did not volunteer for a program, but were eligible; using students in classes that will receive the curriculum (treatment) at a later date; and matching individual characteristics. (Hedrick et al., 1993, p. 59)

Geronimus (1991) provided a good example of creating a strong comparison group. She and her colleagues completed several studies of long-term outcomes for teen mothers. As you are probably aware, the stereotypical outcome for teen mothers is quite dismal: Younger mothers are more likely to have negative experiences such as poverty, high dropout rates, and higher rates of infant mortality. Geronimus believed that family factors, such as socioeconomic status, might be better predictors of these negative outcomes than the actual teen pregnancy. Random assignment for research on this topic would be impossible—you could not randomly assign teenage girls to become pregnant. Quasi-experimentation was thus necessary. In looking for a comparison group that would be as similar as possible, Geronimus decided to use the teenage mothers' sisters who did not become pregnant until later in life. Thus, although the assignment to groups was not random, the groups were presumably very near to equivalence, particularly with respect to family background factors. Interestingly enough, when family background was controlled in this manner, many of the negative outcomes associated with teen pregnancy disappeared. For example, there was no longer any difference in the dropout rates of the two groups. "For indicators of infant health and children's sociocognitive development, at times the trends reversed direction (i.e., controlling for family background, the teen birth group did better than the postponers)" (Geronimus, 1991, p. 465).

In Geronimus's research the "pretest" (actually a matching variable in this case) consisted of finding two women from the same family, one who first became pregnant as a teenager and one who did not get pregnant until after age 20. In this case the groups may still have been nonequivalent, but they were highly equivalent on family background. Sometimes it is impossible to begin with equivalent groups, and the pretest serves much like a baseline measure for comparison with the posttest. In this type of situation the label *nonequivalent groups* seems quite appropriate.

Janet Luehring, a student at Washburn University in Topeka, Kansas, and Joanne Altman, her faculty advisor, used a nonequivalent group design in their research project (Luehring & Altman, 2000). They measured students' performance on the Mental Rotation Task (MRT; Vandenberg & Kuse, 1978). For each item on the MRT, participants saw five three-dimensional shapes, with the first shape being the test stimulus. Two of the other four shapes were matches of the test stimulus when rotated; participants had to identify the two that were the same as the test stimulus. The MRT consists of 20 such items and normally has a 6-minute time limit. The preponderance of evidence from psychological research indicates that men tend to perform better on spatial tasks than women (Luehring & Altman, 2000). Luehring and Altman compared the performance of female students on the MRT to that of male students; the groups, thus, were not equal before the experiment began. The IV in Luehring and Altman's experiment consisted of performing the MRT under timed or untimed conditions. They found that women who performed the MRT under timed conditions made as few errors as men under timed or untimed conditions—only the women under untimed conditions made more errors than the other three groups. Because the two gender groups began the experiment as nonequivalent, the appropriate question after the experiment was not whether a difference existed, but whether the difference was the same as before the experiment (see Figure 13-13A) or whether the difference had changed in some way (see Figure 13-13B). In Luehring and Altman's experiment the difference between the two groups had grown smaller in the timed condition, thus supporting the hypothesis that the IV had an effect on MRT performance for women. Of course, there are several other possible outcomes that would show some effect of the IV. More of Cook and Campbell's (1979) hypothetical outcomes appear in Figure 13-14. Can you interpret each set of findings pictured there?

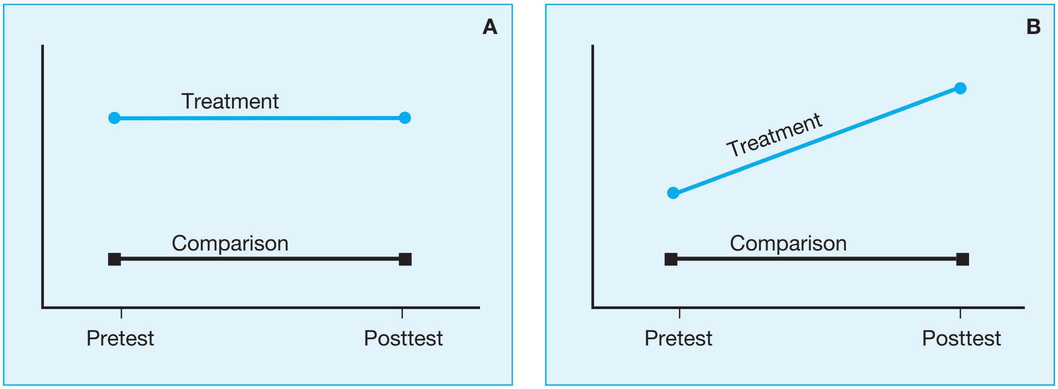


FIGURE 13-13 Two Possible Outcomes in a Nonequivalent Group Design.

Source: Quasi-Experimentation: Design and Analysis, by Thomas D. Cook and Donald T. Campbell, 1979, New York: Houghton-Mifflin. Reprinted by permission.

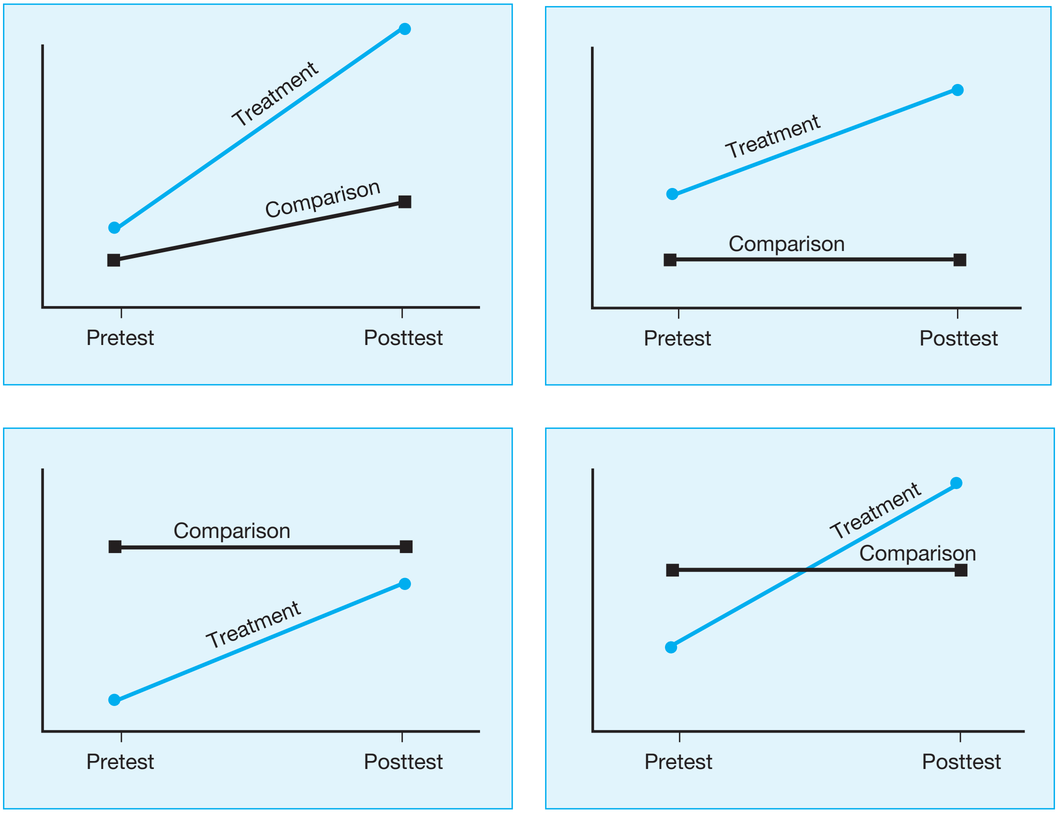


FIGURE 13-14 Several Additional Outcomes in a Nonequivalent Group Design.

Source: Quasi-Experimentation: Design and Analysis, by Thomas D. Cook and Donald T. Campbell, 1979, New York: Houghton-Mifflin and Company. Reprinted by permission.

Thus far, our discussion of this design has seemed similar to that of true experimental designs. What is different about quasi-experimental designs? The most important point to remember is that quasi-experimental designs are more plagued by threats to internal validity. Because you have not used random assignment, your interpretation of the findings must be cautious. Cook and Campbell (1979) isolated four threats to internal validity that are not controlled in the nonequivalent group design. We will list these threats only briefly because they appeared in Chapter 8. First, *maturation* is a potential problem. Because the groups begin as unequal, there is a greater potential that results such as those shown in Figure 13-13B might be due to differential maturation of the groups rather than to the IV. Second, we must consider *instrumentation* in the nonequivalent group design. For example, if we demonstrate nonequivalence of our participants by using a scale during the pretest, we must worry about whether the scale is uniform—are the units of measurement equal throughout the scale? *Statistical regression* is the third internal validity threat present in the nonequivalent group design. Regression is particularly likely to be a problem if we select extreme scorers on the basis of our pretest. Finally, we must consider the threat to internal validity of an *interaction between selection and history*. If some local event differentially affected our treatment and comparison groups, we would have a problem.

In conclusion, the nonequivalent group design is a strong quasi-experimental design. Its strength lies in the fact that "it provides an approximation to the experimental design and that, with care, it can support causal inference" (Hedrick et al., 1993, p. 62). Of course, we must be aware that the threat of confounds is higher than it is in the true experimental designs. Hedrick et al. (1993) warned that "throughout both the planning and execution phases of an applied research project, researchers must keep their eyes open to identify potential rival explanations for their results" (p. 64). Often researchers who use quasi-experimental designs must address potential alternative hypotheses in their research reports.

Interrupted Time-Series Design Another quasi-experimental design, the **interrupted time-series design**, involves measuring a group of participants repeatedly over time (the time series), introducing a treatment (the interruption), and measuring the participants repeatedly again (more of the time series). Look at Figure 13-15 to see a graphic portrayal of an interrupted time-series design. We should make an important point about Figure 13-15: There is nothing magical about using five observations before (O_1–O_5) and after (O_6–O_{10}) the treatment. Any number of observations large enough to establish a pattern can be used (Campbell & Stanley, 1966, showed four before and after; Cook & Campbell, 1979, showed five; Hedrick et al., 1993, showed six before and five after). As you can probably guess, the idea behind an interrupted time-series design is to look for changes in the trend of the data before and after the treatment is applied. Thus, the interrupted time-series design is similar to an A-B design. A change in trend could be shown by a change in the *level* of the behavior (see Figure 13-16A), a change in the *rate* (slope) of the pattern of behavior (see Figure 13-16B), or both (see Figure 13-16C).

Interrupted time-series design A quasi-experimental design, involving a single group of participants, that includes repeated pretreatment measures, an applied treatment, and repeated posttreatment measures.

Interrupted time-series designs have been used for quite some time. Campbell and Stanley (1966) referred to their use in much of the classical research of nineteenth-century biology and physical science. Cook and Campbell (1979) cited a representative 1924 study dealing with the effects of moving from a 10-hour to an 8-hour workday in London. Hedrick and Shipman (1988) used an interrupted time-series design to assess the impact of the 1981 Omnibus Budget Reconciliation Act (OBRA), which tightened eligibility requirements for Aid to Families with Dependent Children (AFDC) assistance. As shown in Figure 13-17, the immediate impact of this legislation was to lessen the number of cases handled by about 200,000; however, the number of cases after the change continued to climb at about the same slope it had before the change. Thus, the tightened eligibility requirements seemed to lower the *level* of the caseload but not its *rate*.

PSYCHO-LOGICAL DETECTIVE

Review the threats to internal validity summarized in Chapter 8. Which threat would seem to create the greatest potential problem for the interrupted time-series design?

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

Figure 13-15 An Interrupted Time-Series Design.

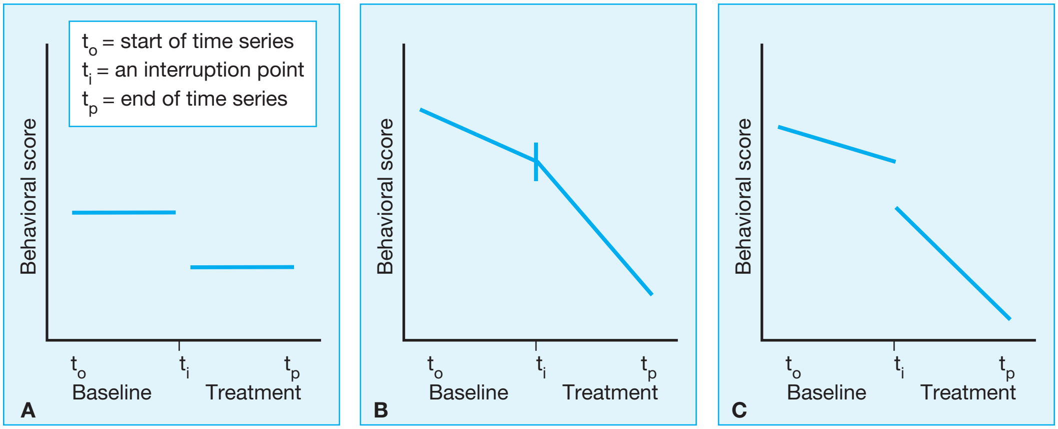


FIGURE 13-16 Potential Changes in Trend in a Time-Series Design. **A**. Change in level, no change in rate. **B**. No change in level, change in rate. **C**. Change in level, change in rate.

Source: Portions of Figure 1 from "Time-Series Analysis in Operant Research," by R. R. Jones, R. S. Vaught, and M. Weinrott, 1977, *Journal of Applied Behavior Analysis*, *10*, pp. 151–166.

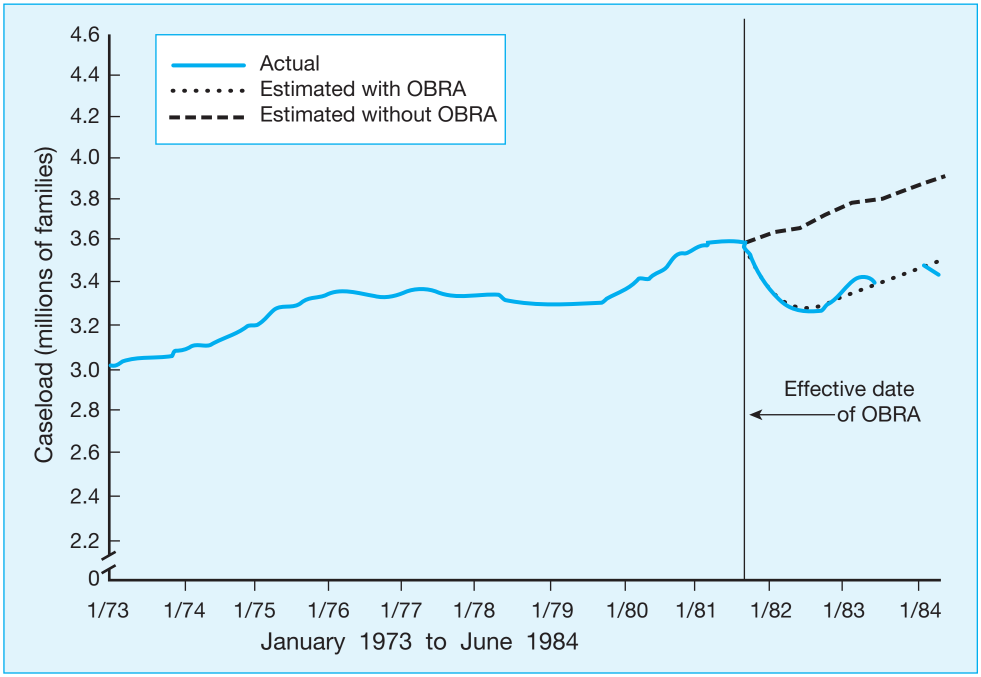


FIGURE 13-17 Impact of Tightened AFDC Requirements on Caseload.

Source: "Multiple Questions Require Multiple Designs: An Evaluation of the 1981 Changes to the AFDC Program," by T. E. Hedrick and S. L. Shipman, *Evaluation Review*, *12*, p. 438. Copyright © 1988 Sage Publications, Inc. Reprinted by permission of Sage Publications, Inc.

According to Cook and Campbell (1979), the main threat to most interrupted time-series designs is *history*. One of the primary features of the interrupted time-series design is the passage of time needed to take many different measurements. This time passage raises the possibility that changes in behavior *could* be due to some important event other than the treatment. Because of the time taken by repeated measurements, another potential threat to internal validity is *maturation*. Repeated pretesting does, however, allow for the assessment of any maturational trends: If scores change at the same rate before and after the treatment, the change is due to maturation. *Instrumentation* could be a problem if record-keeping or scoring procedures change over the course of time. Such a change, of course, would violate the principles of control in any experiment, not just an interrupted time-series design.

Although the interrupted time-series design can control for some of the internal validity threats, we still face the potential problem of history. This threat to internal validity is usually handled in one of three manners. First, Cook and Campbell (1979) advised frequent testing intervals. For example, if you test participants on a weekly rather than monthly, quarterly, or yearly basis, the probability of a major event occurring during the time period between the last pretest and the treatment is low. In addition, if you keep careful records of any possible effect-causing events that occur during the quasi-experiment, it would be a simple matter to discern whether any occurred at the critical period when you administered the treatment. This first approach to controlling history is probably the most widely used because of its ease and the drawbacks involved with the next two solutions.

The next solution to the history threat is to include a comparison (control) group that does not receive the treatment. Such a design appears in Figure 13-18. As you can see, the comparison group receives the same number of measurements at the same times as the treatment (experimental) group. Thus, if any important historical event occurs at the time the experimental group receives the treatment, the comparison group would have the same experience and show the same effect. The only problem with this solution is that the comparison group would most likely be a nonequivalent group because the groups were not randomly assigned. This nonequivalence would put us back in the situation of attempting to control for that difference, with the associated problems we covered in the previous section of this chapter.

The third possible solution to the history problem is probably the best solution, but it is not always possible to do. In essence, this solution involves using an A-B-A format within the interrupted time-series design. The problems, of course, are those that we mentioned earlier in the chapter when dealing with the A-B-A design. Most important, it may not be possible to "undo" the treatment. Once a treatment has been applied, it is not always reversible. Also, if we halt an experiment in the A stage, we are leaving our participants in a nontreatment stage, which may have negative consequences. Hedrick et al. (1993) presented the results of an unintentional interrupted time-series design in an A-B-A format. In 1966 the federal government passed the Highway Safety Act, including a provision that mandated helmets for motorcyclists.

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

FIGURE 13-18 An Interrupted Time-Series Design with Control Group.

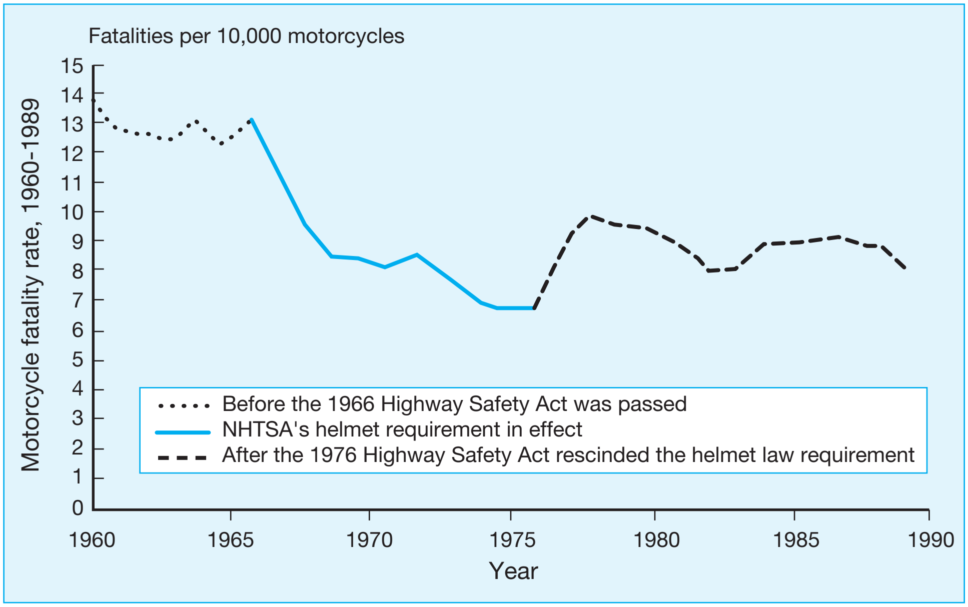


FIGURE 13-19 Effects of Mandatory Helmet Laws and Their Subsequent Repeal on Motorcycle Fatalities.

Source: Motorcycle Helmet Laws Save Lives and Reduce Costs to Society (GAO/RCED-91-170, July), Washington, DC. From *Applied Research Design: A Practical Guide*, by T. E. Hedrick, L. Bickman, and D. J. Rog, 1993, Newbury Park, CA: Sage, p. ??.

In the late 1970s states began to repeal the helmet laws as a result of the pressure being applied by individuals concerned with freedom of choice. If we examine motorcycle fatality rates over many years, we have an A (no restrictions), B (helmet laws), A (fewer restrictions) format for an interrupted time-series design. Figure 13-19 shows a graph presented by Hedrick et al. (1993). Because of the drop in fatalities after the law was passed and the rise in fatalities after the law was repealed in some states, it seems straightforward to derive a cause-and-effect relation from these data. Although this type of design allows for a convincing conclusion, again we must point out that the circumstances that created it are unusual and would be difficult, if not impossible, to recreate in many typical quasi-experimental situations.

In summary, the interrupted time-series design has the ability to uncover cause-and-effect relations. You must be especially careful of history effects when using this design; however, frequent testing can reduce this threat. The interrupted time-series design is particularly helpful when you are dealing with applied types of problems such as therapeutic treatment or in educational settings.

REVIEW SUMMARY

1. **Quasi-experimental designs** are identical to true experimental designs except that participants are not randomly assigned to groups. Thus, our research groups may not be equal before the experiment, which can cause problems in drawing clear conclusions.
2. Unlike the case in ex post facto designs, we are able to control the IV in a quasi-experimental design.

3. There are many situations in which the impossibility of random assignment makes quasi-experimentation necessary.
4. The **nonequivalent group design** involves comparing two groups—one of which receives the IV and a comparison group that does not receive the IV. The groups are nonequivalent because of the lack of random assignment.
5. In the nonequivalent group design, it is imperative to select a comparison group that is as similar as possible to the treatment group.
6. **Maturation, instrumentation, statistical regression**, and **selection–history interactions** are all threats to internal validity in the nonequivalent group design.
7. An **interrupted time-series design** involves measuring participants several times, introducing an IV, and then measuring the participants several more times.
8. **History** is the main threat to internal validity in the interrupted time-series design. It can be controlled by testing frequently, including a comparison group, or removing the treatment after it has been applied (if possible).

Check Your Progress

1. Differentiate between experimental designs, quasi-experimental designs, and ex post facto designs.
2. Give two reasons why you might choose to use a quasi-experimental design rather than an experimental design.
3. Match the design with the appropriate characteristics:

   1. nonequivalent group design
   2. interrupted time-series design

   A. typically has one group of participants
   B. has two groups of participants
   C. involves pretesting participants
   D. does not involve pretesting participants
   E. is prone to the internal validity threat of history
   F. is prone to several internal validity threats
4. What was the key to Geronimus's (1991) research that allowed her to conclude that the effects of teenage pregnancy are not as negative as typically thought?
5. We summarized two interrupted time-series analyses in the text: one dealing with changing AFDC requirements (Figure 13-17) and one dealing with changing motorcycle helmet laws (Figure 13-19). Why are we more certain about our conclusion in the case of the helmet laws than with the AFDC requirements?
6. If Prohibition (the outlawing of alcoholic beverages in the 1920s) were to be treated as an experiment to determine its effects on alcohol consumption, what design would this represent?

   a. nonequivalent group design
   b. single-case design
   c. interrupted time-series design with control group
   d. interrupted time-series design

Key Terms

Internal validity, 300
Confounding variables, 300
Extraneous variables, 300
Cause-and-effect relation, 300
Random assignment, 303
Random selection, 304
Selection, 304
History, 304
Maturation, 304
Testing, 304
Statistical regression, 304
Interactions with selection, 305
Experimental mortality, 305
Instrumentation, 305
Diffusion or imitation of treatments, 305
Single-case experimental design, 308
Case-study approach, 309
Experimental analysis of behavior, 310
Baseline, 312
A, 314
B, 314
A-B design, 314
A-B-A design, 315
A-B-A-B design, 316
Quasi-experimental design, 321
Nonequivalent group design, 323
Interrupted time-series design, 327

Looking Ahead

At this point you may have finished planning, conducting, and analyzing your research project. Still, one task lies ahead—writing the research report, which is the culmination of your research effort. In the next chapter we will cover how researchers write their reports in the American Psychological Association's style.