9 Experimental, Quasi-Experimental, and Ex Post Facto Designs

> Progress is relative: We measure it by noting the degree of change between what was and what is. And we attempt to account for the change by identifying the dynamics that have caused it. Ideally, we must manipulate one possible causal factor while controlling all other possible causal factors; only in this way can we determine whether the manipulated factor has a causal effect on the phenomenon we are studying. To the extent that many potentially causal factors all vary at once, in a confounded manner, we learn little or nothing about what causes what.

In the designs described in preceding chapters, we have made no systematic attempt to determine the causes of the phenomena being studied. But ultimately we often do want to know what leads to what; in other words, we want to identify *cause-and-effect relationships.*

A researcher can most convincingly identify cause-and-effect relationships by using an **experimental design**. In such a design, the researcher considers many possible factors that might cause or influence a particular condition or phenomenon. The researcher then attempts to control for all influential factors *except* those whose possible effects are the focus of investigation.

An example can help clarify the point. Imagine that we have two groups of people. We take steps to make sure that, on average, the two groups are so similar that we can, for all intents and purposes, call them equivalent. We give members of both groups a pretest to measure a particular characteristic in which we are interested—for instance, this might be blood pressure, academic achievement, or purchasing habits. Then we expose only one of the groups to a **treatment** or intervention of some sort—perhaps a new pharmaceutical drug, an instructional method, or an advertising campaign—that we think may have an effect on the characteristic we are studying. Afterward, we give members of both groups a posttest to measure the characteristic once again. If the characteristic changes for the group that received the intervention but does *not* change for the other group, and if everything about the two groups has been the same *except for the intervention,* we can reasonably conclude that the treatment or intervention brought about the change we observed. Because we have not only observed but also *manipulated* the situation, we have used an experimental design.

Some of the research designs we describe in this chapter are true experimental designs; as such, they allow us to identify cause-and-effect relationships. Other designs in this chapter eliminate some—but not all—alternative explanations of an observed change. Yet all of the designs in the chapter have one thing in common: clearly identifiable independent and dependent variables.

We introduced you to independent and dependent variables in Chapter 2, but because these concepts will guide so much of our discussion in this chapter, a brief refresher might be in order here. An **independent variable** is one that the researcher studies as a possible cause of something else. In many of the designs described in this chapter, the researcher directly manipulates the independent variable. In contrast, a **dependent variable** is a variable that is potentially influenced by the independent variable—it is the *"something else"* the independent variable possibly influences, and thus it *depends* on the independent variable. In other words, the hypothesized relationship is this:

Independent variable → Dependent variable

As an example, let's look at a dissertation in educational psychology (Thrailkill, 1996). The researcher wanted to study the effects of three different kinds of lecture material on people's ability to remember information contained in the lecture. Working with undergraduate students, she presented different parts of a lecture on an obscure American Civil War battle in one of three ways: (1) she described certain historical figures and events in such a manner that they were easy to imagine and visualize (*imagery* condition), (2) she included attention-grabbing phrases in the lecture (*attention* condition), or (c) she did neither of these things (*control* condition). In the following examples, the underscored phrases illustrate the modifications made for each of the three conditions; other variations in wording made the three lectures equivalent in length:

> *Imagery:* Lincoln also created the Army of Virginia, incorporating several forces which had been under different commanders. Lincoln set the dimpled, baby-faced young blond Major General John Pope in charge of this new combined force. Being put under his command was objectionable to some of the former commanders. . . .
>
> *Attention:* Lincoln also created the Army of Virginia, incorporating several forces which had been under different commanders. LISTEN TO ME NOW. Lincoln set the less experienced Major General John Pope in charge of this new combined force. Being put under the command of Pope was objectionable to some of the former commanders. . . .
>
> *Control:* Lincoln also created the Army of Virginia, incorporating several forces which had been under different commanders. Lincolw the less experienced junior officer Major General John Pope in charge of this new combined force. Being put under the command of Pope was objectionable to some of the former commanders. (Thrailkill, 1996, p. 62, some underscoring added)

After presenting different parts of the lecture under the three different conditions, the researcher measured the students' recall for the lecture in two ways. She first gave students blank sheets of paper and asked them to write down as much of the lecture as they could remember (a "free recall" task). When they had completed the task, she gave them a multiple-choice test that assessed their memory for specific facts within the lecture. In this study, the independent variable was the nature of the lecture material: easily visualized, attention-getting, or neutral. There were two dependent variables, both of which reflected students' ability to recall facts within the lecture: students' performance on the free recall task and their scores on the multiple-choice test. Thrailkill's hypothesis was confirmed: The students' ability to recall lecture content *depended,* to some extent, on the way in which the content was presented.

The Importance of Control

A particular concern in any experimental study is its **internal validity**, the extent to which its design and the data it yields allow the researcher to draw legitimate conclusions about cause-and-effect and other relationships (see Chapter 4). In experimental designs, internal validity is essential. Without it, a researcher cannot draw firm conclusions about cause and effect—and that is, after all, the whole point of conducting an experimental study.

As an example, suppose we have just learned about a new method of teaching science in elementary school. We want to conduct an experiment to investigate the method's effect on students' science achievement test scores. We find two fifth-grade teachers who are willing to participate in the study. One teacher agrees to use the new method in the coming school year; in fact, she is quite eager to try it. The other teacher wants to continue using the same approach he has always used. Both teachers agree that at the end of the school year we can give their students a science achievement test.

Are the two classes the same in every respect *except for the experimental intervention?* If the students taught with the new method obtain higher science achievement test scores at the end of the year, will we know that the method was the *cause* of the higher scores? The answer to both questions is a resounding *no!* The teachers are different: One is female and the other male, and they almost certainly have different personalities, educational backgrounds, teaching styles, and so on. In addition, the two groups of students may be different; perhaps the students instructed by the new method are, on average, more intelligent or motivated than the other, or perhaps they live in a more affluent school district. Other, more subtle differences may be at work as well, including the interpersonal dynamics in the two classes, and the light, temperature, and noise levels within each classroom. Any of these factors—and perhaps others we haven't thought of—might be reasons for any group differences in achievement test scores we obtain.

Whenever we compare two or more groups that are or might be different in ways *in addition to* the particular treatment or intervention we are studying, we have **confounding variables** in our study. The presence of such variables makes it extremely difficult to draw conclusions about cause-and-effect relationships, because we cannot pin down *what* is the cause of any pattern in the data observed after the intervention. In other words, confounding variables threaten a study's internal validity. In a classic book chapter, Campbell and Stanley (1963) identified several potential threats to the internal validity of an experimental study; we describe them in Figure 9.1.

Controlling for Confounding Variables

To maximize internal validity when a researcher wants to identify cause-and-effect relationships, the researcher needs to control confounding variables in order to rule them out as explanations for any effects observed. Researchers use a variety of strategies to control for confounding variables. Following are several common ones:

1. *Keep some things constant.* When a factor is the *same* for everyone, it cannot possibly account for any differences observed. Oftentimes researchers ensure that different treatments are imposed in the same or very similar environments. They may also seek research participants who share a certain characteristic, such as sex, age, grade level, or socioeconomic status. Keep in mind, however, that restricting the nature of one's sample may lower the *external validity,* or generalizability, of any findings obtained (see Chapter 4's discussion of this concept).

2. *Include a control group.* In Chapter 4 we described a study in which an industrial psychologist begins playing classical music as employees in a typing pool go about their daily task of typing documents. At the end of the month, the psychologist finds that the typists' productivity is 30% higher than it was during the preceding month. The increase in productivity may or may not be due to the classical music. There are too many possible confounding variables—personnel changes, nature of the documents being typed, numbers of people out sick or on vacation during the two-month period, even just the knowledge that an experiment is being conducted—that may also account for the typists' increased productivity.

To better control for such extraneous variables, researchers frequently include a **control group**, a group that receives either no intervention or a "neutral" intervention that should have little or no effect on the dependent variable. The researchers then compare the performance of this group to an **experimental group**—also known as a **treatment group**—that participates in an intervention.

As you should recall from Chapter 4, people sometimes show improved performance simply because they know they are participating in a research study—a phenomenon known as *reactivity* and, more specifically, the *Hawthorne effect.* To take this fact into account, a researcher sometimes gives the people in a control group a **placebo** that has the appearance of having an effect but in reality *shouldn't* have an effect. For instance, a researcher studying the effects of a new arthritis medication might give some participants a particular dosage of the medicine and give others a

FIGURE 9.1

Potential threats to the internal validity in an experimental study

When a researcher studies the possible effects of an intervention on some other (dependent) variable, a number of confounding variables can come into play that threaten the study's internal validity and thereby also jeopardize any cause-and-effect conclusions the researcher might hope to draw. Campbell and Stanley (1963) have identified the following potential threats to internal validity, which can be present either singly or in combination:

1. *History:* An uncontrolled outside event occurring between two measurements of the dependent variable brings about a change in the dependent variable. For example, a noteworthy event in the local community might change participants' knowledge, abilities, or emotional states in ways that affect the second measurement of the dependent variable.
2. *Maturation:* A change in participants' characteristics or abilities might simply be the result of the passage of time. For example, children might make normal developmental gains in eye-hand coordination or intellectual ability.
3. *Testing:* Taking a test at one time influences participants' performance during a subsequent administration of the test, perhaps simply as a result of practice in taking the test. For example, people who take a multiple-choice test at one time may gain general test-taking skills that enhance their performance on a subsequent multiple-choice test.
4. *Instrumentation:* A change occurs in how a measurement instrument is used from one time to the next. For example, a researcher might have one research assistant rate participants' performance on the first occasion but have a different research assistant judge their performance on the subsequent occasion. Any observed change might be the result of the two assistants' differing standards for rating the performance. (This threat to internal validity reflects a problem with interrater reliability; see Chapter 4).
5. *Statistical regression:* People who score extremely high or low on a measure at one time are likely to score in a less extreme manner on the same measure at a later time; that is, extreme scorers tend to "drift" toward more average performance during a subsequent measure. For example, a researcher might assign people to one of two groups—"high-anxiety" or "low-anxiety"—based on their extremely high or low scores on a self-report questionnaire designed to measure general anxiety level. Especially if the initially extreme scores were the result of people's temporary circumstances—circumstances that might make them feel either exceptionally anxious or, instead, quite "mellow" on the first testing—the supposedly high-anxiety people would become less anxious and the supposedly low-anxiety people would become more anxious regardless of any experimental interventions the two groups might undergo.
6. *Selection:* A bias exists in how members of different groups in a study are chosen. For example, when recruiting college students for a study, a researcher might put all students enrolled in an 8:00 a.m. class in one treatment group and all students enrolled in a 2:00 p.m. class in another treatment group. Students taking the early-morning class might be different in some significant way from those taking the afternoon class (e.g., the sleeping habits of the two groups might be different).
7. *Attrition:** Members of different groups drop out of the study at proportionally different rates. For example, one group in a study might lose 25% of its members before the final measurement, whereas another group might lose only 5% of its members. Thus, even if the two groups were equivalent with regard to important characteristics at the beginning of the study, they might be different in some significant way later in the study simply as a result of the differential dropout rate.

Campbell and Stanley listed an eighth threat to internal validity as well: an *interaction* among two of the threats listed above. For example, if students in an 8:00 a.m. class are assigned to one treatment group and students in a 2:00 p.m. class are assigned to a different treatment group, and if students in the 8:00 a.m. group drop out of the study in greater numbers than students in the 2:00 p.m. group, any final differences observed in the dependent variable might be the result of the fact that early risers are, for some reason, more likely to drop out than students who like to sleep in a bit. In this situation, it becomes virtually impossible to disentangle possible effects of an experimental intervention from effects of (a) the selection bias, (b) the differing dropout rates, and (c) the interaction of these two confounding variables.

**Note:* Campbell and Stanley use the term *experimental mortality* for this threat to internal validity, but the term *attrition* is more commonly seen in contemporary research literature.

similar-looking sugar pill. Or a researcher investigating a new approach to treating test anxiety might use the new treatment with some individuals but give other individuals general relaxation training that, although possibly beneficial in other ways, won't necessarily address their test anxiety.

We must emphasize—and we emphasize it quite strongly—that any researcher who incorporates placebos in a study must consider *three ethical issues* related to the use of placebos. First is the principle of informed consent: Participants in the study must be told that the study includes a placebo treatment as well as an experimental treatment and that they won't know which treatment they have received until the study has ended. Second, if participants in the study have actively sought help for a medical, psychological, or other significant problem, those who initially receive the placebo treatment should, at the conclusion of the study, be given the opportunity to receive more effective treatment. (This is assuming, of course, that the treatment *is* more effective than the placebo.) Third, and most important, when studying a treatment related to life-threatening situations (e.g., a new drug for terminal cancer, a new psychotherapeutic technique for suicidal teenagers), the researcher must seriously weigh (a) the benefits of the new knowledge that can be gained by a control group receiving no treatment against (b) the lives that may be saved by including all participants in the treatment group.

Our last point raises an issue we cannot possibly resolve for you here. Should you find yourself having to make a decision about the best research design to use in a life-and-death situation, you should consult with your professional colleagues, the internal review board at your institution, and, of course, your own conscience.

3. *Randomly assign people to groups.* In Chapter 8 we spoke at length of the value of selecting people at random to participate in a descriptive research study; such random selection enhances the probability that any results obtained for the sample also apply to the population from which the sample has been drawn. In experimental studies, researchers use random selection for a different purpose: to assign participants within their sample to various groups.

In any research study involving human beings or other living things, members of the sample are likely to be different from one another in many ways that are relevant to the variables under investigation. For example, earlier in the chapter we described a situation in which a researcher wants to compare two methods of teaching elementary school science. The students in the study will almost certainly differ from one another in intelligence, motivation, educational opportunities at home, and other factors that will affect their performance on the achievement test given at the end of the school year. It would be virtually impossible to control for such variables by having all students in the study have the *same* intelligence, the *same* motivation, the *same* kinds of outside opportunities, and so on.

As an alternative to keeping some characteristics the same for everyone, a researcher can, instead, randomly assign participants to groups. When people have been selected for one group or another on a random basis, the researcher can reasonably assume that, *on average, the groups are quite similar* and that *any differences between them are due entirely to chance.* In fact, many inferential statistical tests—especially those that allow the researcher to make comparisons among two or more groups—are based on the assumption that group membership is randomly determined and that any pretreatment differences between the groups result from chance alone.

4. *Assess equivalence before the treatment with one or more pretests.* Sometimes random assignment to two different groups simply isn't possible; for instance, researchers may have to study groups that already exist (e.g., students in classrooms, participants in different medical treatment programs). An alternative in this situation is to assess other variables that might influence the dependent variable and determine whether the groups are similar with respect to those variables. If the groups *are* similar, the probability that such variables could account for any future group differences is reduced considerably.

Another strategy is to identify **matched pairs**: pairs of people—one in each of two groups being compared—who are identical or very similar with respect to characteristics that may potentially have an effect on the dependent varaible. For instance, a researcher comparing the achievement test scores of students in two different instructional programs might identify pairs of students of the same sex and age who have similar IQ scores. A researcher comparing two different treatments for a particular illness might match patients according to sex, age, and duration and intensity of the illness. In either case, the researcher does not study the data collected for *all* people in the two groups, only the people who are part of "matched sets" that he or she has identified. A researcher who uses this approach will, in the final research report, explain in what way(s) the participants in the study have been matched. For example, he or she might say, "Pairs were matched on the basis of sex, age, and IQ."

One problem with assessing before-treatment equivalence with pretests is that the researcher rules out *only the variables that he or she has actually assessed and determined to be equivalent across groups.* The design does not rule other influential factors that the researcher has not assessed and perhaps not even considered.

5. *Expose participants to all experimental conditions.* Still another strategy for controlling for individual differences is to *use participants as their own controls*—that is, to have every participant in the study undergo all experimental and control treatments and then assess the effects of each treatment independently. Any independent variable that is systematically varied for every participant is known as a **within-subjects variable**, and an approach that includes a within-subjects variable is known as a **within-subjects design**. You may also see the term **repeated-measures design** used in reference to this approach.

As an example, let's return to the dissertation involving three different lecture methods and their possible effects on recall for lecture content (Thrailkill, 1996). The researcher's sample consisted of volunteer students who were enrolled in three sections of an undergraduate class in educational psychology, and she planned to give the lecture just three times, once to each class. The lecture was about an American Civil War battle sufficiently obscure that participants were unlikely to have had any prior knowledge about it; thus, participants' prior knowledge about the battle was a constant—they all had *zero* prior knowledge—rather than a confounding variable. The researcher divided the lecture into three parts of approximately equal length and wrote three versions of each part, one version each for the imagery, attention, and control conditions. She combined the three versions of the three lecture parts such that each class received the different treatments in a different sequence, as follows:

	PART OF LECTURE		
	First Part	*Middle Part*	*Last Part*
Group 1	Attention	Imagery	Control
Group 2	Control	Attention	Imagery
Group 3	Imagery	Control	Attention

In this manner, all participants in her study were exposed to the two treatments and the control condition, and each condition occurred in all possible places (first, second, and third) in the sequence.

In the study just described, the researcher used a within-subjects variable (type of intervention: imagery vs. attention vs. control) to compensate for the fact that participants had not been randomly assigned to the three class sections in her sample. Sometimes researchers use a similar strategy with just a single group, and in some cases with just a single individual. You will learn some strategies for showing causation in single-group and single-individual studies later in the chapter, when we explore *quasi-experimental designs.*

6. *Statistically control for confounding variables.* Sometimes researchers can control for known confounding variables, at least in part, through statistical techniques. Such techniques as *partial correlation, analysis of covariance* (ANCOVA), and *structural equation modeling* are suitable for

this purpose. We briefly describe each of these in Chapter 11. Should you choose to use one of them in your own research, we urge you to consult one or more statistics books for guidance about their use and appropriateness for various research situations.

Keep in mind, however, that statistically controlling confounding variables is no substitute for controlling for them in one's research design if at all possible. *A carefully controlled experimental design is the only approach that allows you to draw firm conclusions about cause-and-effect relationships.*

Overview of Experimental, Quasi-Experimental, and Ex Post Facto Designs

In true experimental research, the researcher manipulates the independent variable and examines its effects on another, dependent variable. A variety of research designs have emerged that differ in the extent to which the researcher manipulates the independent variable and controls for confounding variables—in other words, the designs differ in the degree to which they have *internal validity.* In the upcoming sections, we present a number of possible designs, which we have divided into five general categories: *pre-experimental designs, true experimental designs, quasi-experimental designs, ex post facto designs,* and *factorial designs.* Altogether we describe 16 different designs that illustrate various ways—some more effective than others—of attempting to identify cause-and-effect relationships. Some of our discussion is based on designs identified by Campbell and Stanley (1963).[1]

We illustrate the designs using tables that have this general format:

Group	Time →	
Group 1		
Group 2		

Each group in a design is shown in a separate row, and the things that happen to the group over time are shown in separate cells within the row. The cells have one of four notations:

Tx: Indicates that a *treatment* (reflecting the independent variable) is presented.

Obs: Indicates that an *observation* (reflecting the dependent variable) is made.

—: Indicates that nothing occurs during a particular time period.

Exp: Indicates a previous *experience* (an independent variable) that some participants have had and others have not; the experience has *not* been one that the researcher could control.

The nature of these tables will become more apparent as we proceed.

As you read about the 16 designs, keep in mind that they are hardly an exhaustive list; researchers can modify or combine them in various ways. For example, although we will be limiting ourselves to studies with only one or two groups (perhaps one treatment group and one control group), it is entirely possible to have two or more treatment groups (each of which is exposed to a different variation of the independent variable) and, in some cases, two control groups (perhaps one getting a placebo and another getting no intervention at all). More generally, the designs we describe here should simply provide a starting point that gets you thinking about how you might best tackle your own research problem.

[1] In particular, Designs 1–6 and Designs 8–11 are based on those that Campbell and Stanley described. However, when describing Design 11, we use the contemporary term *reversal time-series design* rather than Campbell and Stanley's original term *equivalent time-samples design.*

Pre-Experimental Designs

In **pre-experimental designs**, it isn't possible to show cause-and-effect relationships, because either (a) the independent "variable" doesn't vary or (b) experimental and control groups are not comprised of equivalent or randomly selected individuals. Such designs are helpful only for forming tentative hypotheses that should be followed up with more controlled studies.

Design 1: One-Shot Experimental Case Study

The one-shot experimental case study is probably the most primitive type of experiment that might conceivably be termed "research." An experimental treatment (Tx) is introduced, and then a measurement (Obs)—a posttest of some sort—is administered to determine the effects of the treatment. This design is shown in the following table:

Group	Time →	
Group 1	Tx	Obs

The design has low internal validity because it is impossible to determine whether participants' performance on the posttest is the result of the experimental treatment per se. Many other variables may have influenced participants' performance, such as physiological maturation or experiences elsewhere in the participants' general environment. Perhaps the characteristic or behavior observed after the treatment existed *before* the treatment as well. The reality is that with a single measurement or observation, we have no way of knowing whether the situation has changed or not, let alone whether it has changed as a result of the intervention.

One-shot experimental case studies may be at the root of many common misconceptions. For example, imagine that we see a boy sitting on the damp ground in mid-April. The next day he has a sore throat and a cold. We conclude that sitting on the damp earth caused him to catch cold. Thus, the design of our "research" thinking is something like this:

Exposure to cold, damp ground (Tx) → Child has a cold (Obs)

Such "research" may also "support" such superstitious folk beliefs as these: If you walk under a ladder, you will have bad luck; Friday the 13th is a day of catastrophes; a horseshoe above the door brings good fortune to the house. Someone observed an event, then observed a subsequent event, and linked the two together as cause and effect.

Be careful not to confuse the one-shot experimental case study method with the case study design of many qualitative studies. As described in Chapter 6, case study research involves extensive engagement in a research setting—a far cry from basing conclusions on a single observation.

Although the one-shot experimental case study is simple to carry out, its results are, for all intents and purposes, meaningless. At the very least, researchers should use the design described next.

Design 2: One-Group Pretest–Posttest Design

In a one-group pretest–posttest design, a single group (a) undergoes a pre-experimental observation or evaluation, then (b) is administered the experimental treatment, and finally (c) is observed or evaluated again after the treatment. This design can be represented as follows:

Group	Time →		
Group 1	Obs	Tx	Obs

Suppose an elementary school teacher wants to know if simultaneously reading a story and listening to it on audiotape will improve the reading skills of students in his class. He gives his students a standardized reading test, then has them simultaneously read and listen to simple stories every day for eight weeks, and then administers an alternate form of the same standardized reading test. If the students' test scores improve over the eight-week period, the teacher might conclude—perhaps accurately, but perhaps not—that the simultaneous-reading-and-listening intervention was the cause of the improvement.

Now suppose an agronomist hybridizes two strains of corn. She finds that the hybrid strain is more disease-resistant and has a better yield than either of the two parent types. She concludes that the hybridization process has made the difference. Once again we have an Obs–Tx–Obs design: The agronomist measures the disease level of the parent strains (Obs), then develops a hybrid of the two strains (Tx), and then measures the disease level of the next generation (Obs).

In a one-group pretest–posttest design, we at least know that a change has taken place. However, we have not ruled out other possible explanations for the change. In the case of the elementary school teacher's study, improvement in reading scores may have been due to other activities within the classroom curriculum, to more practice taking the reading test, or simply to the fact that the students were eight weeks older. In the case of the agronomist's experiment, changes in rainfall, temperature, or soil conditions may have been the primary reason for the healthier corn crop.

Design 3: Static Group Comparison

The static group comparison involves both an experimental group and a control group. Its design takes the following form:

Group	Time →	
Group 1	Tx	Obs
Group 2	—	Obs

An experimental group is exposed to a particular experimental treatment; the control group is not. After the treatment, both groups are observed and their performance compared. In this design, however, no attempt is made to obtain equivalent groups or even to examine the groups to determine whether they are similar before the treatment. Thus, we have no way of knowing if the treatment actually causes any observed differences between the groups.

The three designs just described leave much to be desired in terms of drawing conclusions about what causes what. The experimental designs we describe next are far superior in this respect.

True Experimental Designs

In contrast with the three very simple designs just described, **experimental designs** offer a greater degree of control and, as a result, greater internal validity. The first three of the four designs we discuss in this section share one thing in common: People or other units of study are *randomly assigned to groups.* Such random assignment guarantees that any differences between the groups are probably quite small and, in any case, are due entirely to chance. The last design in this section involves a different strategy: presenting all treatments and any control conditions to a single group.

Design 4: Pretest–Posttest Control Group Design

In a pretest–posttest control group design, an experimental group and a control group are carefully selected through appropriate randomization procedures. The experimental group is observed, subjected to the experimental treatment, and observed again. The control group is isolated from any influences of the experimental treatment; it is simply observed both at the beginning and at the end of the experiment. The basic format for the pretest–posttest control group design is as follows:

	Group	Time →		
Random Assignment	Group 1	Obs	Tx	Obs
	Group 2	Obs	—	Obs

Such a design, simple as it is, solves two major problems associated with pre-experimental designs. We can (a) determine whether a change takes place after the treatment, and, if so, we can (b) eliminate most other possible explanations (in the form of confounding variables) as to why the change has taken place. Thus, we have a reasonable basis on which to draw a conclusion about a cause-and-effect relationship.

Design 5: Solomon Four-Group Design

One potential problem in the preceding design is that the process of observing or assessing people before administering the experimental treatment may, in and of itself, influence how people respond to the treatment. For instance, perhaps the pretest increases people's motivation: It makes them want to benefit from the treatment they receive. Such an effect is another instance of the *reactivity* effect described in Chapter 4.

To address the question, What effect does pretesting have?, Solomon (1949) proposed an extension of the pretest–posttest control group design that involves four groups, as depicted in the following table:

	Group	Time →		
Random Assignment	Group 1	Obs	Tx	Obs
	Group 2	Obs	—	Obs
	Group 3	—	Tx	Obs
	Group 4	—	—	Obs

The addition of two groups that are not pretested provides a distinct advantage. If the researcher finds that in the final observation, Groups 3 and 4 differ in much the same way that Groups 1 and 2 do, then the researcher can more easily generalize his or her findings to situations in which no pretest has been given. In other words, the Solomon four-group design enhances the *external validity* of the study.

Compared to Design 4, this design obviously involves a larger sample and demands more of the researcher's time and energy. Its principal value is in eliminating pretest influence; when such elimination is desirable, the design is ideal.

Design 6: Posttest-Only Control Group Design

Some life situations defy pretesting. You cannot pretest the forces in a thunderstorm or a hurricane, nor can you pretest growing crops. Additionally, at times you may be unable to locate a suitable pretest, or, as just noted, the very act of pretesting can influence the results of the experimental manipulation. In such circumstances, the posttest-only control group design offers a possible solution. The design may be thought of as the last two groups of the Solomon four-group design. The paradigm for the posttest-only approach is as follows:

	Group	Time →	
Random Assignment	Group 1	Tx	Obs
	Group 2	—	Obs

Random assignment to groups is, of course, critical in the posttest-only design. Without it, the researcher has nothing more than a static group comparison (Design 3), from which, for reasons previously noted, the researcher has a difficult time drawing inferences about cause and effect.

Design 7: Within-Subjects Design

Earlier we introduced you to the nature of a within-subjects design—also known as a repeated-measures design—in which all participants receive all treatments (including any control conditions) in a research study. Note that we have switched from the term *participant* to the term *subject* here. The latter term has a broader meaning than *participants* in that it can be used to refer to a wide variety of populations—perhaps human beings, dogs, pigeons, or laboratory rats.

In a good within-subjects design, the various treatments are administered very close together in time, in some cases simultaneously. If we use the subscripts *a* and *b* to designate the different treatments and treatment-specific measures, then in its simplest form a within-subjects design is as follows:

Group	Time →	
Group 1	Tx_a	Obs_a
	Tx_b	Obs_b

As an example, imagine that a researcher wants to study the effects of illustrations in teaching science concepts to sixth graders. The researcher creates a short textbook that presents, say, 20 different concepts. In the text, all 20 concepts are defined and described with similar precision and depth. In addition, the text illustrates 10 of those concepts (chosen randomly) with pictures or diagrams. After students read the text, they take a quiz that assesses their understanding of the 20 concepts, and the researcher computes separate quiz scores for the illustrated and nonillustrated concepts. If the students perform better on quiz items for illustrated concepts than on items for nonillustrated ones, the researcher can reasonably conclude that, yes, illustrations help students learn science more effectively. In other words, the researcher has identified a cause-and-effect relationship: Illustrations improve science learning.

For a within-subjects design to work, the various forms of treatment must be such that their effects are fairly localized and unlikely to "spread" beyond specifically targeted behaviors. This is the case in the study just described: The illustrations help students learn the particular concepts that have been illustrated but don't help students learn science more generally. In contrast, it would not make sense to use a within-subjects design to study the effects of two different psychotherapeutic techniques to reduce adolescents' criminal behaviors: If the same group of adolescents receives both treatments and then shows a significant reduction in juvenile offenses, we might suspect that either treatment could have had a fairly broad impact.

Ideally, too, the two different treatments should be administered repeatedly, one after another, in a balanced but somewhat random order. For example, in the textbook that presents both illustrated and nonillustrated science concepts, we might begin with an illustrated concept, then have two nonillustrated ones, then another illustrated one, another nonillustrated one, two illustrated ones, and so on, with the presentation of the two conditions being evenly balanced throughout the book.

With the last point in mind, let's return to the dissertation involving the American Civil War lecture described earlier. Each group received each of the three treatments: the imagery, attention, and control conditions. The logistics of the study were such that it was difficult to intermingle the three treatments throughout the lecture; instead, the researcher administered first one treatment (e.g., attention), then another (e.g., imagery), and finally the third (e.g., control). Had the researcher limited her study to a single group, she could not have ruled out an alternative explanation—*when* in the lecture the information appeared (whether it appeared near the beginning, in the middle, or at the end)—for the results she obtained. By using three different groups, each of which had any particular condition in a different part of the lecture, she was able to eliminate that alternative explanation. Strictly speaking, however, because the researcher could neither randomize assignment to groups nor randomly distribute different treatment conditions throughout the lecture, her study is probably better characterized as a quasi-experimental study than a true experimental study. We look more closely at quasi-experimental designs now.

Quasi-Experimental Designs

In the preceding discussion of true experimental designs, we emphasized the importance of *randomness,* either in the selection of group members in a multiple-groups study or in the presentation of different treatments in a single-group study. Sometimes, however, randomness is either impossible or impractical. In such situations, researchers often use **quasi-experimental designs.** When they conduct quasi-experimental studies, they don't control for all confounding variables and so cannot completely rule out some alternative explanations for the results they obtain. They must take whatever variables and explanations they have not controlled for into consideration when they interpret their data.

Design 8: Nonrandomized Control Group Pretest–Posttest Design

The nonrandomized control group pretest–posttest design can perhaps best be described as a compromise between the static group comparison (Design 3) and the pretest–posttest control group design (Design 4). Like Design 3, it involves two groups to which participants have not been randomly assigned. But it incorporates the pretreatment observations of Design 4. In sum, the nonrandomized control group pretest–posttest design can be depicted as follows:

Group	Time →		
Group 1	Obs	Tx	Obs
Group 2	Obs	—	Obs

Without random assignment, there is, of course, no guarantee that the two groups are similar in every respect prior to the experimental treatment or intervention—no guarantee that any differences between them are due entirely to chance. However, an initial observation (e.g., a pretest) can confirm that the two groups are at least similar in terms of the dependent variable under investigation. If, after one group has received the experimental treatment, we then find group differences with respect to the dependent variable, we might reasonably conclude that the posttreatment differences are probably the result of that treatment.

Identifying matched pairs in the two groups is one way of strengthening the pretest–posttest control group design. For instance, if we are studying the effect of a particular preschool program on children's IQ scores, we might find pairs of children—each pair including one child who is enrolled in the preschool program and one who is not—who are the same sex and age and have similar IQ scores before the program begins. Although we cannot rule out all other possible explanations in this situation (e.g., it may be that the parents who enroll their children in the preschool program are, in general, more concerned about their children's cognitive development), we can at least rule out *some* alternative explanations.

Design 9: Simple Time-Series Design

In its simplest form, a time-series design consists of making a series of observations (i.e., measuring the dependent variable on several occasions), introducing an intervention or other new dynamic into the system, and then making additional observations. If a substantial change is observed in the second series of observations in comparison to the first series, we might reasonably conclude that the cause of the change was the factor introduced into the system. This design thus looks something like the following:

Group	Time →								
Group 1	Obs	Obs	Obs	Obs	Tx	Obs	Obs	Obs	Obs

In such studies, the sequence of observations made prior to the treatment is typically referred to as **baseline** data.

Such a design has been widely used in the physical and biological sciences. Sir Alexander Fleming's discovery that *Penicillium notatum* (a mold) could inhibit staphylococci (a type of bacteria) is an example of this type of design. Fleming had been observing the growth of staphylococci on a culture plate. Then, unexpectedly, a culture plate containing well-developed colonies of staphylococci was contaminated with the spores of *Penicillium notatum.* Fleming observed that the bacteria near the mold seemed to disappear. He intentionally repeated the situation: After periodically observing the bacteria, he introduced the mold. Each time he used this procedure, his subsequent observations were the same: no staph germs near the mold.

The major weakness of this design is the possibility that some other, unrecognized event in the laboratory or outside world may occur at approximately the same time that the experimental treatment does, reflecting the *history* factor described in Figure 9.1. If this other event is actually the cause of the change, any conclusion that the treatment has brought about the change will, of course, be an erroneous one.

Design 10: Control Group, Time-Series Design

In a variation of the time-series design, two groups are observed over a period of time, but one group (a control) doesn't receive the experimental treatment. The design is configured as follows:

Group	Time →								
Group 1	Obs	Obs	Obs	Obs	Tx	Obs	Obs	Obs	Obs
Group 2	Obs	Obs	Obs	Obs	—	Obs	Obs	Obs	Obs

This design has greater internal validity than the simple time-series design (Design 8). If an outside event is the cause of any changes we observe, then presumably the performance of *both* groups will be altered after the experimental treatment takes place. If, instead, the experimental treatment is the factor that affects performance, then we should see a change only for Group 1.

Design 11: Reversal Time-Series Design

The reversal time-series design uses a within-subjects approach as a way of minimizing—though not entirely eliminating—the probability that outside effects might bring about any changes observed. The intervening experimental treatment is sometimes present, sometimes absent, and we measure the dependent variable at regular intervals. Thus, we have the following design:

Group	Time →							
Group 1	Tx	Obs	—	Obs	Tx	Obs	—	Obs

To illustrate, suppose we are interested in whether audiovisual materials help students learn astronomy. On some days we might include audiovisual materials in a lesson, and on other days we might omit them. We can then measure how effectively students learn under both conditions. If the audiovisual materials do, in fact, promote student learning, we should see consistently better student performance on those days.

Design 12: Alternating Treatments Design

A variation on the reversal time-series design involves including two or more different forms of experimental treatment in the design. Referring to the two different forms of treatment with the notations Tx_a and Tx_b, we can depict this design in the following way:

Group	Time →													
Group 1	Tx_a	Obs	—	Obs	Tx_b	Obs	—	Obs	Tx_a	Obs	—	Obs	Tx_b	Obs

If such a sequence were pursued over a long enough time span, we would hope to see different effects for the two different treatments.

Design 13: Multiple Baseline Design

Designs 11 and 12 are based on the assumption that the effects of any single treatment are temporary and limited to the immediate circumstances. Thus, these designs won't work if a treatment is apt to have long-lasting and perhaps fairly general effects. Furthermore, if an experimental treatment is likely to be quite beneficial for all participants, then ethical considerations may discourage us from including an untreated control group. In such instances, a multiple baseline design provides a good alternative. This design requires at least two groups. Prior to the treatment, baseline data are collected for all groups, and then the treatment itself is introduced at a different time for each group. In its simplest form, a multiple baseline design might be configured as follows:

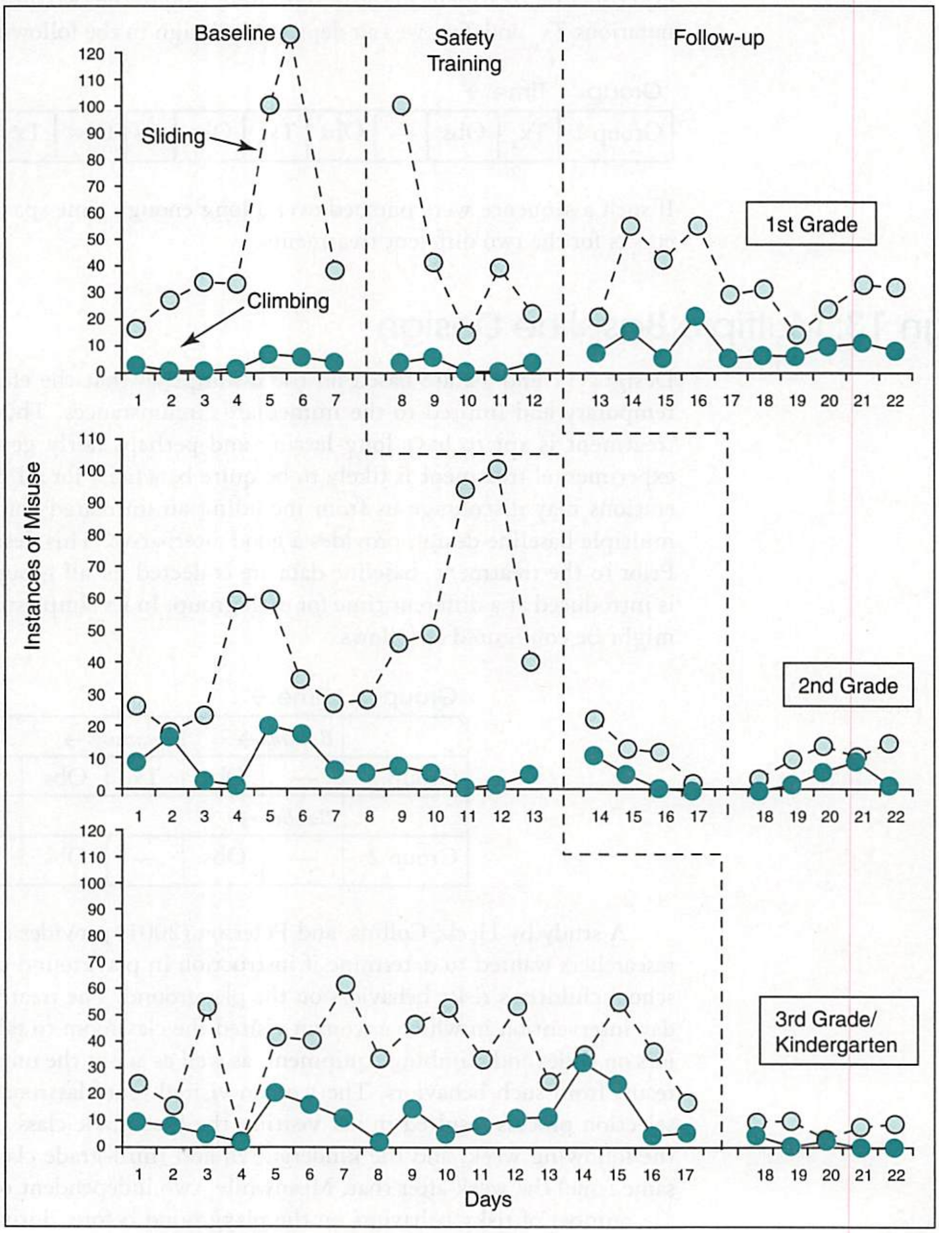
Group	Time →					
	Baseline →		*Treatment →*			
Group 1	—	Obs	Tx	Obs	Tx	Obs
	Baseline →				*Treatment →*	
Group 2	—	Obs	—	Obs	Tx	Obs

A study by Heck, Collins, and Peterson (2001) provides an example of this approach. The researchers wanted to determine if instruction in playground safety would decrease elementary school children's risky behaviors on the playground. The treatment in this case involved a five-day intervention in which a woman visited the classroom to talk about potentially risky behaviors on slides and climbing equipment, as well as about the unpleasant consequences that might result from such behaviors. The woman visited four classrooms on different weeks; a random selection process resulted in her visiting the first-grade class one week, the second-grade class the following week, and the kindergarten and third-grade classes (which went to recess at the same time) the week after that. Meanwhile, two independent observers simultaneously counted the number of risky behaviors on the playground before, during, and after the interventions in the four classrooms. The data they collected are depicted in Figure 9.2; numbers of risky behaviors on the slide are shown with the lighter dots, whereas those on the climbing equipment are shown with the darker dots. Notice that each group has data for three time periods: a pre-intervention baseline period, the five-day safety-training period, and a posttraining follow-up period. As you can see, once safety training began, the children in the second-grade and third-grade/kindergarten groups showed noticeable declines in risky behaviors on the slide and, to a lesser extent, on the climbing equipment (where risky behavior was relatively infrequently to begin with). Because the behavior changes occurred at different times for these two groups, and in particular because the changes for each group occurred at about the time that the group began its safety training, the researchers reasonably concluded that the training itself (rather than some other factor) was probably the reason for the changes. The first graders, who received the training first, showed little or no benefit from it, especially for the climbing equipment. Perhaps the trainer was still perfecting her training procedures that first week; however, we have no way of knowing for sure why the training appeared to be relatively ineffective for the first group.

FIGURE 9.2

Instances of risky behavior on slides and climbers by grade level; third graders and kindergartners shared a single recess

Reprinted from "Decreasing Children's Risk Taking on the Playground" by A. Heck, J. Collins, and L. Peterson, 2001, *Journal of Applied Behavior Analysis, 34,* p. 351. Reprinted with permission of the Society for the Experimental Analysis of Behavior, Inc.



Using Designs 11, 12, and 13 in Single-Subject Studies

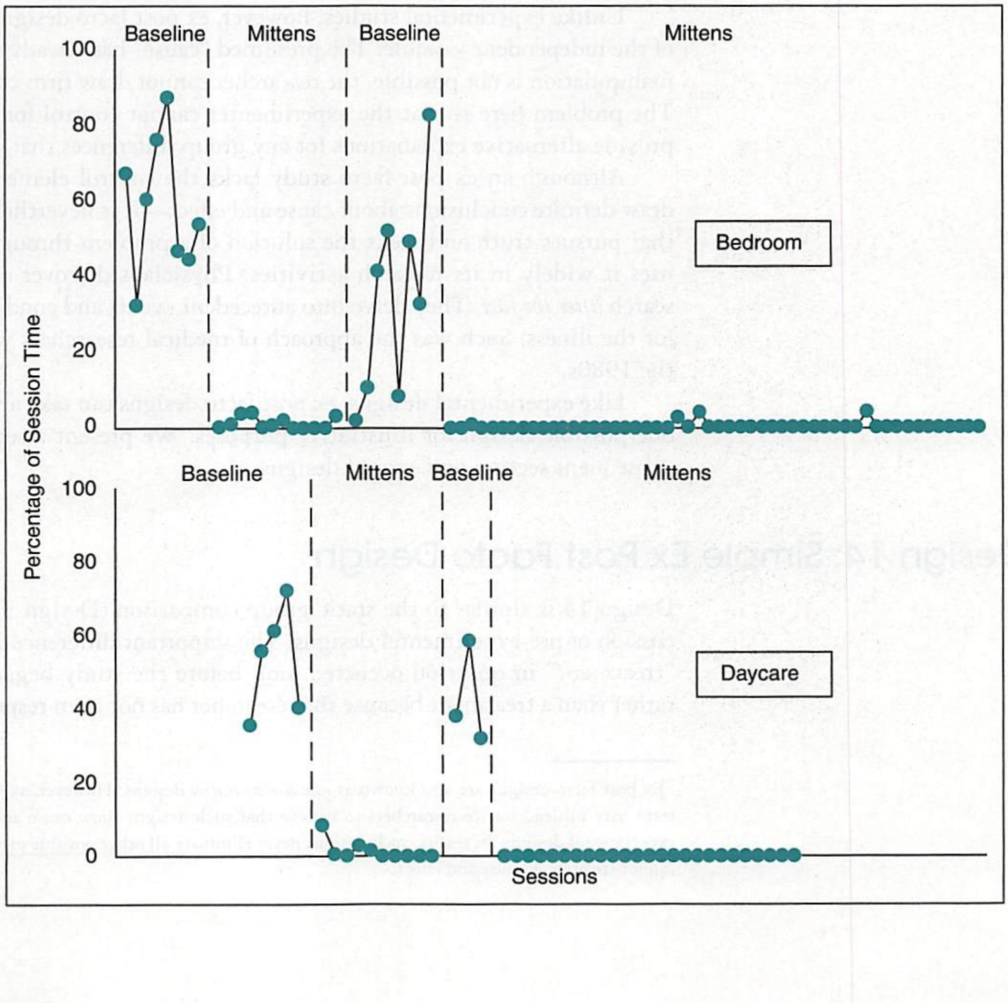
Reversal, alternating treatments, and multiple baseline designs can be used not only with groups but also with single individuals, in what are collectively known as **single-subject designs**. A study by Deaver, Miltenberger, and Stricker (2001) illustrates how a researcher might use two of these—reversal and multiple baseline—simultaneously. A 2-year-old girl named Tina had been referred for treatment because she often twirled her hair with her fingers so vigorously that she pulled out some of her hair. On one occasion she wrapped her hair around a finger so tightly that the finger began to turn blue and the hair had to be removed with scissors. Tina engaged in such behavior primarily when she was alone (e.g., at naptime); hence, there was no parent or other adult present to discourage it. The researchers identified a simple treatment—putting thin cotton mittens on Tina's hands—and wanted to document its effect. They videotaped Tina's behaviors when she was lying down for a nap in either of two settings, her bedroom at home or her daycare center, and two observers independently counted the number of hair twirling incidents as they watched the videotapes. Initially, the observers collected baseline data. Then, during separate time periods for the bedroom and daycare settings, they gave Tina the mittens to wear during naptime. After reversing back to baseline in both settings, they had Tina wear the mittens once again. The percentages of time that Tina twirled her hair in the two settings over the course of the study are presented in Figure 9.3.

In both the bedroom and daycare settings, the researchers alternated between baseline and treatment; this is the *reversal* aspect of the study. Furthermore, they initiated and then later reinstituted the treatment at different times in the two settings; this is the *multiple baseline* aspect of the study. Figure 9.3 consistently shows dramatic differences in hair twirling during baseline versus mittens conditions, leading us to conclude that the mittens, rather than some other factor, were almost certainly the reason for the disappearance of hair twirling.

FIGURE 9.3

Percentage of session time in which hair twirling was observed both in the bedroom and at daycare

Reprinted from "Functional Analysis and Treatment of Hair Twirling in a Young Child" by C. M. Deaver, R. G. Miltenberger, & J. M. Stricker, 2001, *Journal of Applied Behavior Analysis, 34,* p. 537. Reprinted with permission of the Society for the Experimental Analysis of Behavior, Inc.


Baseline
Mittens
Baseline
Mittens
Bedroom
Baseline
Mittens
Baseline
Mittens
Daycare
Percentage of Session Time
Sessions