OUTLINE

Experimental designs are often touted as the most rigorous of all research designs, or as the gold standard against which all other designs are judged. In one sense, they probably are. If you can implement an experimental design well (and that is a big *if* indeed), the experiment is probably the strongest design with respect to **internal validity** (see Chapter 7).

This chapter introduces the idea of an experimental design and describes why it is strong in internal validity. I show that the key distinguishing feature of experimental design—random assignment to group—depends on the idea of **probabilistic equivalence** and explain what that means. I then try to head off one of the biggest sources of confusion to most students—the distinction between **random selection** and **random assignment**. Then I get into the heart of the chapter, describing how to classify the different experimental designs, presenting each type in turn.

internal validity
The approximate truth of inferences regarding cause-effect or causal relationships.

probabilistic equivalence
The notion that two groups, if measured infinitely, would on average perform identically. Note that two groups that are probabilistically equivalent would seldom obtain the exact same average score in a real setting.

random selection
Process or procedure that assures that the different units in your population are selected by chance.

random assignment
Process of assigning your sample into two or more subgroups by chance. Procedures for random assignment can vary from flipping a coin to using a table of random numbers to using the random number capability built into a computer.

9-1 Introduction to Experimental Design

9-1a Experimental Designs and Internal Validity

As mentioned previously, experimental designs are usually considered the strongest of all designs in internal validity (see the discussion on internal validity in Chapter 7). Why? Recall that internal validity is at the center of all **causal** or cause-effect inferences. When you want to determine whether some program or treatment causes some outcome or outcomes to occur, you are interested in having strong internal validity. Essentially, you want to assess the proposition:

causal
Pertaining to a cause-effect relationship.

If *X*, then *Y*.

Or, in more colloquial terms:

If the program is given, then the outcome occurs.

Unfortunately, it's not enough to show that when the program or treatment occurs, the expected outcome also happens because many reasons, other than the program, might account for why you observed the outcome. To show that there is a causal relationship, you have to simultaneously address the two propositions:

If *X*, then *Y*.
and
If not *X*, then not *Y*.

Or, once again more colloquially:

If the program is given, then the outcome occurs.
and
If the program is *not* given, then the outcome does *not* occur.

If you are able to provide evidence for both of these propositions, you've in effect isolated the program from all of the other potential causes of the outcome. You've shown that when the program is present, the outcome occurs, and when it's not present, the outcome doesn't occur. That points to the causal effectiveness of the program.

Think of all this like a fork in the road. Down one path, you implement the program and observe the outcome. Down the other path, you don't implement the program and the outcome doesn't occur. But, can you take both paths in the road in the same study? How can you be in two places at once? Ideally, what you want is to have the same conditions—the same people, context, time, and so on—and see whether when the program is given you get the outcome and when the program is not given you don't. Obviously, you can never achieve this hypothetical situation. If you give the program to a group of people, you can't simultaneously not give it! So, how do you get out of this apparent dilemma?

Perhaps you just need to think about the problem a little differently. What if you could create two groups or contexts that are as similar as you can possibly make them? If you could be confident that the two situations are comparable, you could administer your program in one (and see whether the outcome occurs) and not give the program in the other (and see whether the outcome doesn't occur). If the two contexts are comparable, this is like taking both forks in the road simultaneously. You can have your cake and eat it too, so to speak.

That's exactly what an experimental design tries to achieve. In the simplest type of experiment, you create two groups that are equivalent to each other. One group (the program or treatment group) gets the program, and the other group (the comparison or **control group**) does not. In all other respects, the groups are treated the same. They have similar people, live in similar contexts, have similar backgrounds, and so on. Now, if you observe differences in outcomes between these two groups, the differences must be due to the only thing that differs between them—that one received the program and the other didn't.

control group
A group, comparable to the program group, that did not receive the program.

Okay, so how do you create two equivalent groups? The approach used in experimental design is to assign people randomly from a common pool of people into the two groups. The experiment relies on this idea of random assignment to groups as the basis for obtaining two similar groups. Then, you give one the program or treatment and you don't give it to the other. You observe the same outcomes in both groups.

The key to the success of the experiment is in the random assignment. In fact, even with random assignment, you never expect the groups you create to be exactly the same. How could they be, when they are made up of different people? You rely on the idea of probability and assume that the two groups are probabilistically equivalent, or equivalent within known probabilistic ranges.

If you randomly assign people to two groups, and you have enough people in your study to achieve the desired probabilistic equivalence, you can consider the experiment strong in internal validity and you probably have a good shot at assessing whether the program causes the outcome(s). (See the discussion of statistical power and sample size in Chapter 11.)

However, many things can go wrong. You may not have a large enough sample. Some people might refuse to participate in your study or drop out part way through. You might be challenged successfully on ethical grounds. (After all, to use this approach you have to deny the program to some people who might be equally deserving of it as others.) You might meet resistance from the staff members in your study who would like some of their favorite people to get the program.

The bottom line here is that experimental design is intrusive and difficult to carry out in most real-world contexts, and because an experiment is often an intrusion, you are setting up an artificial situation so that you can assess your causal relationship with high internal validity. As a result, you are limiting the degree to which you can generalize your results to real contexts where you haven't set up an

FIGURE 9–1 Notation for the basic two-group, posttest-only, randomized experimental design

R	X	O
R		O

external validity
The degree to which the conclusions in your study would hold for other persons in other places and at other times.

two-group, posttest-only, randomized experiment
A research design in which two randomly assigned groups participate. Only one group receives the program, and both groups receive a posttest.

single-group threats to internal validity
A threat to internal validity that occurs in a study that uses only a single program or treatment group and no comparison or control.

multiple-group threats
An internal validity threat that occurs in studies that use multiple groups—for instance, a program and a comparison group.

selection mortality
A threat to internal validity that arises when there is differential nonrandom dropout between groups during the study.

selection-testing
A threat to internal validity that occurs when a differential effect of taking the pretest exists between groups on the posttest.

selection-instrumentation threats
A threat to internal validity that results from differential changes in the test used for each group from pretest to posttest.

social threats to internal validity
Threats to internal validity that arise because social research is conducted in real-world human contexts where people will react to not only what affects them, but also to what is happening to others around them.

experiment. That is, you have reduced your **external validity** to achieve greater internal validity.

In the end, there is no simple answer (no matter what anyone tells you). If the situation is right, an experiment is a strong design, but it isn't automatically so.

Experimental design is a complex subject in its own right. I've been discussing the simplest of experimental designs—a two-group program versus comparison-group design; but there are many experimental design variations that attempt to accomplish different things or solve different problems. In this chapter, you'll explore the basic experimental design, look at the major variations, and learn the principles that underlie all experimental-design strategies.

9-1b Two-Group Experimental Designs

The simplest of all experimental designs is the **two-group, posttest-only, randomized experiment** (see Figure 9–1). In design notation, it has two lines—one for each group—with an *R* at the beginning of each line to indicate that the groups were randomly assigned.

One group gets the treatment or program (the *X*) and the other group is the comparison group and doesn't get the program. (Note that you could alternatively have the comparison group receive the standard or typical treatment, in which case this study would be a relative comparison.)

Notice that a pretest is not required for this design. Usually you include a pretest to determine whether groups are comparable prior to the program. However, because this design uses random assignment, you can assume that the two groups are at least probabilistically equivalent to begin with and the pretest is not required (although you'll see with covariance designs later in this chapter that a pretest may still be desirable in this context).

In this design, you are most interested in determining whether the two groups are different after the program. Typically, you measure the groups on one or more measures (the *O*s in the notation) and you compare them by testing for the differences between the means using a *t*-test or one-way analysis of variance (ANOVA), which is covered in Chapter 11.

The posttest-only randomized experiment is the strongest of all research designs with respect to the threats to internal validity as shown in Figure 9–2. The figure indicates (with a checkmark) that it is strong against all the **single-group threats to internal validity** because it's not a single-group design! (Tricky, huh?) It's also strong against the **multiple-group threats** except for **selection mortality**. For instance, it's strong against the **selection-testing** and **selection-instrumentation threats** because it doesn't use repeated measurement. The selection-mortality threat can be a problem if there are differential rates of dropouts in the two groups. This could result if the treatment or program is a noxious or negative one (such as a painful medical procedure like chemotherapy) or if the control group condition is painful or intolerable. This design is susceptible to all of the **social threats to internal validity**. Because the design requires random assignment in some institutional settings such as schools, it is more likely to utilize persons who would be aware of each other and of the conditions to which you have assigned them.

The posttest-only, randomized experimental design is, despite its simple structure, one of the best research designs for assessing cause-effect relationships. It is

FIGURE 9–2 Threats to internal validity for the posttest-only, randomized experimental design

history	✓
maturation	✓
testing	✓
instrumentation	✓
mortality	✓
regression to the mean	✓
selection	✓
selection - history	✓
selection - maturation	✓
selection - testing	✓
selection - instrumentation	✓
selection - mortality	✗
selection - regression	✓
diffusion or imitation	✗
compensatory equalization	✗
compensatory rivalry	✗
resentful demoralization	✗

relatively easy to execute, and because it uses only a posttest, it is relatively inexpensive. However, there are many variations on this simple experimental design. You can begin to explore these by looking at how you classify the various experimental designs (see Section 9-2, Classifying Experimental Designs).

would use several groups to control for the [illegible] having an effect.

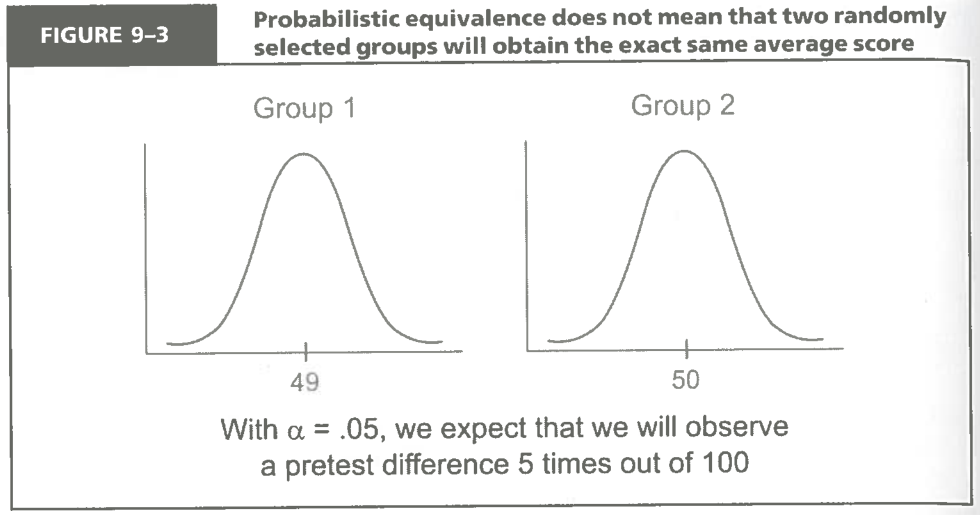
9-1c Probabilistic Equivalence

What do I mean by the term *probabilistic equivalence*, and why is it important to experimental design? Well, to begin with, I certainly *don't* mean that two groups are equal to each other. When you deal with human beings, it is impossible to ever say that any two individuals or groups are equal or equivalent. Clearly the important term in the phrase is *probabilistic*. This means that the type of equivalence you have is based on the notion of probabilities. In more concrete terms, probabilistic equivalence means that you know *perfectly* the odds of finding a difference between two groups. Notice, it doesn't mean that the means of the two groups will be equal. It just means that you know the odds that they won't be equal. Figure 9–3 shows two groups, one with a mean of 49 and the other with a mean of 51. Could these two groups be probabilistically equivalent even though their averages are different? Certainly!

You achieve probabilistic equivalence through the mechanism of random assignment to groups. When you randomly assign to groups, you can calculate the chance that the two groups will differ just because of the random assignment (that is, by chance alone). Let's say you are assigning a group of first-grade students to two groups. Furthermore, let's assume that the average test scores for these children for a standardized test with a population mean of 50 were 49 and 51, respectively. You might conduct a *t*-test to see whether the means of the two randomly assigned groups are statistically different. Through random assignment and the law of large numbers, the chance that they will be different is 5 out of 100 when you set the significance level to .05 (that is, $p = .05$). In other words, 5 times out of every 100, when you randomly assign two groups, you can expect to get a significant difference at the .05 level of significance.

When you assign randomly, groups can differ only due to chance assignment because their assignment is entirely based on the randomness of assignment. If, by chance, the groups differ on one variable, you have no reason to believe that they

FIGURE 9–3 **Probabilistic equivalence does not mean that two randomly selected groups will obtain the exact same average score**



will automatically be different on any other. Even if you find that the groups differ on a pretest, you have no reason to suspect that they will differ on a posttest. Why? Because their pretest difference had to be a chance one. So, when you randomly assign, you are able to assume that the groups do have a form of equivalence. You don't expect them to be equal; but you can expect them to be probabilistically equal.

9-1d Random Selection and Assignment

Random selection is how you draw the sample of people for your study from a population. *Random assignment* is how you assign the sample that you draw to different groups or treatments in your study.

It is possible to have both random selection and assignment in a study. Let's say you drew a random sample of 100 clients from a population list of 1000 current clients of your organization. That is **random sampling**. Now, let's say you randomly assign 50 of these clients to get some new additional treatment and the other 50 to be controls. That's **random assignment**.

random sampling
Process or procedure that assures that different units in your population are selected into a sample by chance.

random assignment
Process of assigning your sample into two or more subgroups by chance. Procedures for random assignment can vary from flipping a coin to using a table of random numbers to using the random number capability built into a computer.

It is also possible to have only one of these (random selection or random assignment) but not the other in a study. For instance, if you do not randomly draw the 100 cases from your list of 1000 but instead just take the first 100 on the list, you do not have random selection. You could, however, still randomly assign this nonrandom sample to treatment versus control. Or, you could randomly select 100 from your list of 1000 and then nonrandomly (haphazardly) assign them to treatment or control groups.

It's also possible to have neither random selection nor random assignment. In a typical nonequivalent-groups design (see Chapter 10) in education you might nonrandomly choose two fifth-grade classes to be in your study. This is nonrandom selection. Then, you could arbitrarily assign one group to get the new educational program and the other to be the control group. This is nonrandom (or nonequivalent) assignment.

Random selection is related to sampling (see Chapter 2). Therefore it is most closely related to the external validity (or generalizability) of your results. After all, researchers randomly sample so that their research participants better represent the larger group from which they're drawn. Random assignment is most closely related to design. In fact, when you randomly assign participants to treatments you have, by definition, an experimental design. Therefore, random assignment is most related to internal validity (see Chapter 7). After all, researchers randomly assign to help ensure that their treatment groups are similar to each other (equivalent) prior to the treatment.

9-2 Classifying Experimental Designs

Although many experimental design variations exist, you can classify and organize them using a simple signal-to-noise ratio metaphor. In this metaphor, assume that what you observe or see in a research study can be divided into two components: the signal and the noise. (By the way, this is directly analogous to the discussion of signal and noise in the true score theory of measurement discussed in Chapter 3.) Figure 9–4 shows a time series with a slightly downward slope. However, because there is so much variability or noise in the series, it is difficult even to detect the downward slope. When you divide the series into its two components, you can clearly see the slope.

In most research, the signal is related to the key variable of interest—the construct you're trying to measure or the program or treatment that's being implemented. The noise consists of all of the random factors in the situation that make it harder to see the signal: the lighting in the room, local distractions, how people felt that day, and so on. You can construct a ratio of these two by dividing the signal by the noise (see Figure 9–5). In research, you want the signal to be high relative to the noise. For instance, if you have a powerful treatment or program (meaning a strong signal) and good measurement (that is, low noise), you have a better chance of seeing the effect of the program than if you have either a strong program and weak measurement or a weak program and strong measurement.

You can further classify the experimental designs into two categories: signal enhancers or noise reducers. Doing either of these things—enhancing signal or reducing noise—improves the quality of the research. The *signal-enhancing experimental designs* are called the **factorial designs**. In these designs, the focus is almost entirely on the setup of the program or treatment, its components, and its major dimensions. In a typical factorial design, you would examine several different varia-

factorial designs
Designs that focus on the program or treatment, its components, and its major dimensions and enable you to determine whether the program has an effect, whether different subcomponents are effective, and whether there are interactions in the effects caused by subcomponents.

FIGURE 9–4 **How an observed time series can be decomposed into its signal and noise**

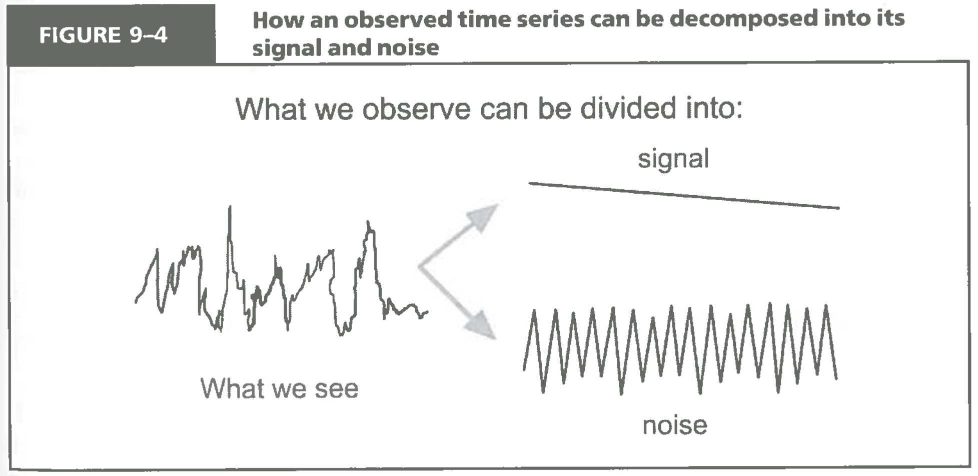


FIGURE 9–5 **The signal-to-noise ratio is simply a fraction where signal is divided by noise**

$$\frac{\text{signal}}{\text{noise}}$$

tions of a treatment. Factorial designs are discussed in the next section of this chapter.

The two major types of *noise-reducing experimental designs* are covariance designs and blocking designs. In these designs, you typically use information about the makeup of the sample or about pre-program variables to remove some of the noise in your study. Covariance and blocking designs are discussed in Section 9-3.

9-3 Factorial Designs

Factorial designs focus on the signal in your research by directly manipulating your program or features of your program or treatment. Factorial designs are especially efficient because they enable you to examine which features or combinations of features of your program or treatment have an effect. I'll start with the simplest factorial design, show you why it is such an efficient approach, explain how to interpret the results, and then move on to more advanced variations.

9-3a The Basic 2 × 2 Factorial Design

Probably the easiest way to begin understanding factorial designs is by looking at an example (see Figure 9–6). Imagine a design where you have an educational program in which you would like to look at a variety of program variations to see which works best. For instance, you would like to vary the amount of time the children receive instruction, with one group getting 1 hour of instruction per week and another getting 4 hours per week. In addition, you'd like to vary the setting so that one group gets the instruction in class (probably pulled off into a corner of the classroom) and the other group is pulled out of the classroom for instruction in another room. You could think about having four separate studies to do this, but when you vary the amount of time in instruction, which setting would you use: in class or pull out? And, when you study setting, what amount of instruction time would you use: 1 hour, 4 hours, or something else?

With factorial designs, you don't have to compromise when answering these questions. You can have it both ways if you cross each of your two times in instruction conditions with each of your two settings. Let's begin by doing some defining of terms. In factorial designs, a factor is a major independent variable. This example has two factors: time in instruction and setting. A **level** is a subdivision of a factor. In this example, time in instruction has two levels and setting has two levels.

level
In an experimental design, a subdivision of a factor into components or features.

Sometimes you depict a factorial design with a numbering notation. In this example, you can say that you have a 2 × 2 (spoken two-by-two) factorial design. In this notation, the *number of numbers* tells you how many factors there are and the *number values* tell you how many levels. A 3 × 4 factorial design has two factors, where one factor has three levels and the other has four. The order of the numbers makes no difference and you could just as easily term this a 4 × 3 factorial design. You can easily determine the number of different treatment groups that you have in any factorial design by multiplying through the number notation. For instance, the school study example has 2 × 2 = 4 groups. A 3 × 4 factorial design requires 3 × 4 = 12 groups.

FIGURE 9–6 An example of a basic 2 × 2 factorial design

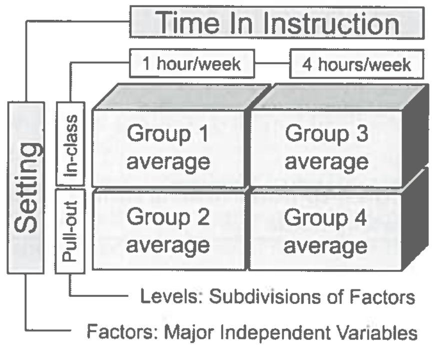


FIGURE 9–7 Design notation for a 2 × 2 factorial design

R X_{11} O
R X_{12} O
R X_{21} O
R X_{22} O

You can also depict a factorial design in design notation. Because of the treatment-level combinations, it is useful to use subscripts on the treatment (*X*) symbol. Figure 9–7 shows that there are four groups, one for each combination of levels of factors. It also shows that the groups were randomly assigned and that this is a posttest-only design.

Now, let's look at a variety of different results you might get from this simple 2 × 2 factorial design. Each of the following figures describes a different possible outcome. Each outcome is shown in table form (the 2 × 2 table with the row and column averages) and in graphic form (with each factor taking a turn on the horizontal axis). Take the time to understand how and why the information in the tables agrees with the information in both of the graphs. Also study the graphs and figures to verify that the pair of graphs in each figure show the exact same information graphed in two different ways. The lines in the graphs are technically not necessary; they are a visual aid that enables you to track where the averages for a single level go across levels of another factor. Keep in mind that the values in the tables and graphs are group averages on the outcome variable of interest. In this example, the outcome might be a test of achievement in the subject being taught. Assume that scores on this test range from 1 to 10, with higher values indicating greater achievement. You should study carefully the outcomes in each figure to understand the differences between these cases.

The Null Outcome The **null case** is a situation in which the treatments have no éffect. Figure 9–8a assumes that even if you didn't give the training, you would expect students to score a 5 on average on the outcome test. You can see in this hypothetical case that all four groups score an average of 5 and therefore the row and column averages must be 5. You can't see the lines for both levels in the graphs because one line falls right on top of the other.

null case
A situation in which the treatment has no effect.

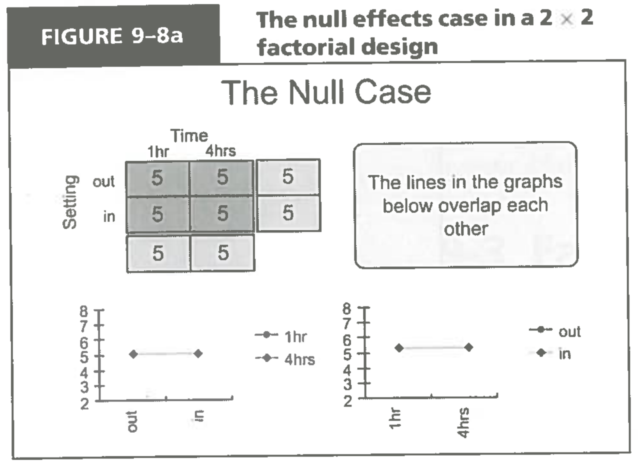
The Main Effects A **main effect** is an outcome that is a consistent difference between levels of a factor. For instance, you would say there's a main effect for setting if you find a statistical difference between the averages for the in-class and pull-out groups, *at all levels* of time in instruction. Figure 9–8b depicts a main effect of time. For all settings, the 4-hour/week condition worked better than the 1-hour/week condition. It is also possible to have a main effect for setting (and none for time).

main effect
An outcome that shows consistent differences between all levels of a factor.

In the second main effect graph, shown in Figure 9–8c, you see that in-class training was better than pull-out training for all amounts of time.

Finally, it is possible to have a main effect on both variables simultaneously, as depicted in the third main effect (Figure 9–8d). In this instance, 4 hours/week

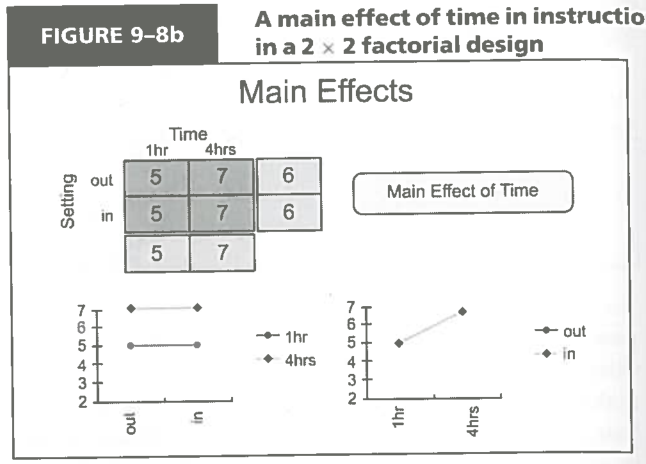
FIGURE 9–8a The null effects case in a 2 × 2 factorial design



Setting	1hr	4hrs	
out	5	5	5
in	5	5	5
	5	5	

The lines in the graphs below overlap each other

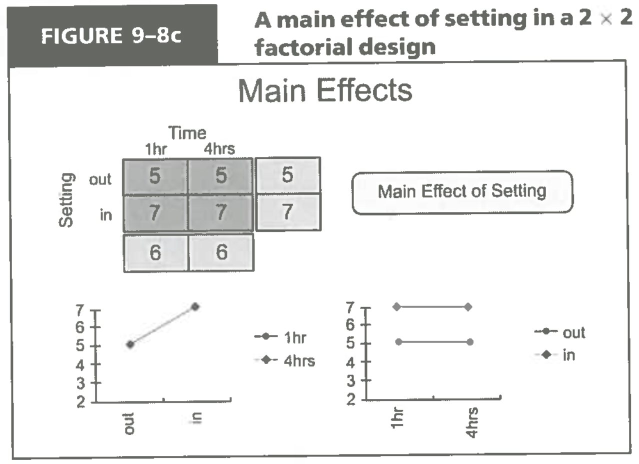
FIGURE 9–8b A main effect of time in instruction in a 2 × 2 factorial design



Setting	1hr	4hrs	
out	5	7	6
in	5	7	6
	5	7	

Main Effect of Time

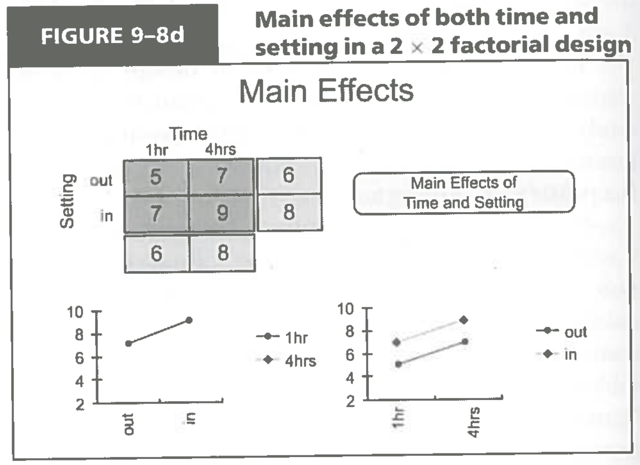
FIGURE 9–8c A main effect of setting in a 2 × 2 factorial design



Setting	1hr	4hrs	
out	5	5	5
in	7	7	7
	6	6	

Main Effect of Setting

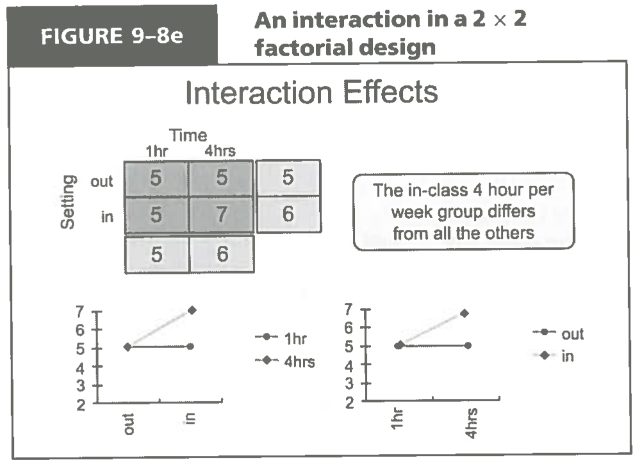
FIGURE 9–8d Main effects of both time and setting in a 2 × 2 factorial design



Setting	1hr	4hrs	
out	5	7	6
in	7	9	8
	6	8	

Main Effects of Time and Setting

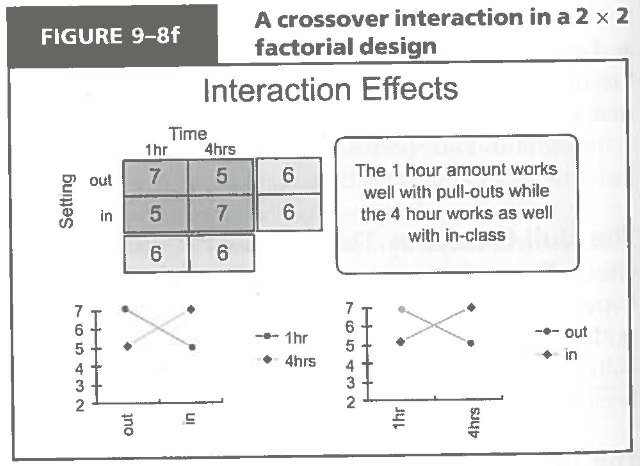
FIGURE 9–8e An interaction in a 2 × 2 factorial design



Setting	1hr	4hrs	
out	5	5	5
in	5	7	6
	5	6	

The in-class 4 hour per week group differs from all the others

FIGURE 9–8f A crossover interaction in a 2 × 2 factorial design



Setting	1hr	4hrs	
out	7	5	6
in	5	7	6
	6	6	

The 1 hour amount works well with pull-outs while the 4 hour works as well with in-class

always works better than 1 hour/week and in-class setting always works better than the pull-out setting.

Interaction Effects If you could look at only main effects, factorial designs would be useful. But, because of the way you combine levels in factorial designs, they also enable you to examine the **interaction effects** that exist between factors. An interaction effect exists when differences on one factor depend on which level you are on in another factor. It's important to recognize that an interaction is between factors, not levels. You wouldn't say there's an interaction between four hours/week and in-class treatment. Instead, you would say that there's an interaction between time and setting, and then you would describe the specific levels involved.

interaction effect
An effect that occurs when differences on one factor depend on which level you are on another factor.

How do you know whether there is an interaction in a factorial design? There are three ways you can determine whether an interaction exists. First, when you run the statistical analysis, the statistical table will report on all main effects and interactions. Second, you know there's an interaction when you can't talk about an effect on one factor without mentioning the other factor. If you can say at the end of your study that time in instruction makes a difference, you know that you have a main effect and not an interaction (because you did not have to mention the setting factor when describing the results for time). On the other hand, when you have an interaction, it is impossible to describe your results accurately without mentioning both factors. Finally, you can always spot an interaction in the graphs of group means; whenever lines are not parallel, an interaction is present! If you check out the main effect graphs in Figure 9–8c, you will notice that all of the lines within a graph are parallel. In contrast, for all of the interaction graphs, you will see that the lines are not parallel.

In the first interaction effect graph (Figure 9–8e), one combination of levels—4 hours/week and in-class setting—shows better results than the other three.

The second interaction (see Figure 9–8f) shows more complex crossover interaction. Here, at 1 hour/week the pull-out group does better than the in-class group whereas at 4 hours/week the reverse is true. Furthermore, both of these combinations of levels do equally well.

Factorial design has several important features. First, it gives you great flexibility for exploring or enhancing the signal (treatment) in your studies. Whenever you are interested in examining treatment variations, factorial designs should be strong candidates as the designs of choice. Second, factorial designs are efficient. Instead of conducting a series of independent studies, you are effectively able to combine these studies into one. Finally, factorial designs are the only effective way to examine interaction effects.

So far, you have only looked at a simple 2 × 2 factorial design structure. You may want to look at some factorial design variations in the following section to get a deeper understanding of how these designs work. You may also want to examine how to approach the statistical analysis of factorial experimental designs (see Chapter 1).

9-3b Factorial Design Variations

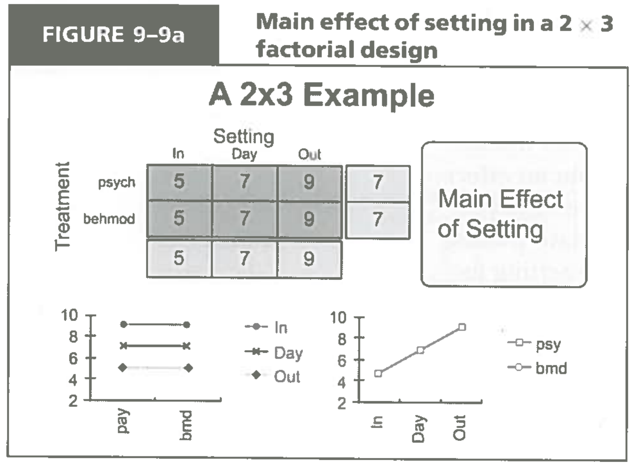
This section discusses a number of different factorial designs. I'll begin with a two-factor design where one of the factors has more than two levels. Then I'll introduce the three-factor design. Finally, I'll present the idea of the incomplete factorial design.

A 2 × 3 Example For these design presentations, I'll construct a hypothetical study designed to assess the effect of different treatment combinations for cocaine abuse. Here, the dependent measure is a severity-of-illness rating performed by the treatment staff. The outcome ranges from 1 to 10, where higher scores indicate more severe illness: in this case, more severe cocaine addiction. Furthermore, assume that the levels of treatment are as follows:

- Factor 1: Treatment
  - Psychotherapy
  - Behavior modification
- Factor 2: Setting
  - Inpatient
  - Day treatment
  - Outpatient

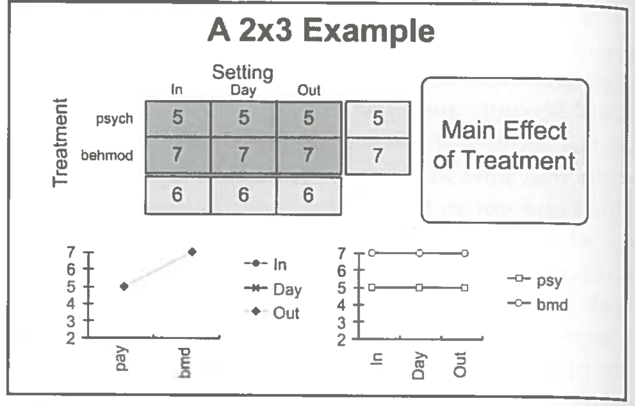
Note that the setting factor in this example has three levels.

FIGURE 9–9a Main effect of setting in a 2 × 3 factorial design



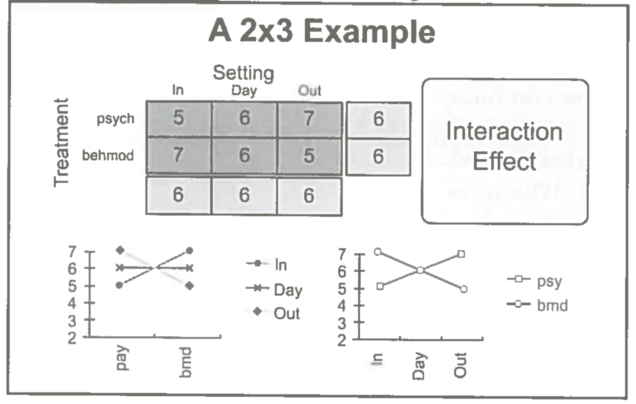
Treatment	In	Day	Out	
psych	5	7	9	7
behmod	5	7	9	7
	5	7	9	

FIGURE 9–9b Main effect of treatment in a 2 × 3 factorial design



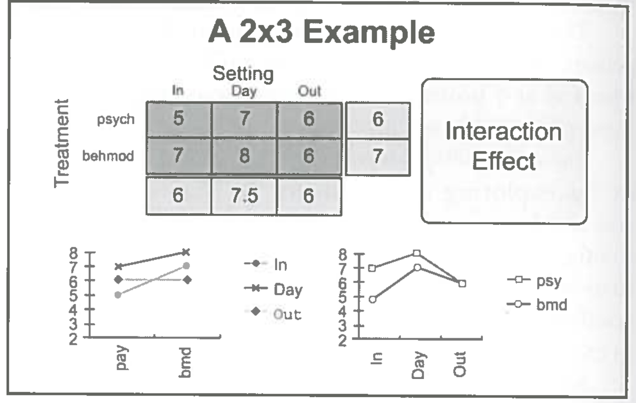
Treatment	In	Day	Out	
psych	5	5	5	5
behmod	7	7	7	7
	6	6	6	

FIGURE 9–9c An interaction effect in a 2 × 3 factorial design



Treatment	In	Day	Out	
psych	5	6	7	6
behmod	7	6	5	6
	6	6	6	

FIGURE 9–9d An interaction effect in a 2 × 3 factorial design



Treatment	In	Day	Out	
psych	5	7	6	6
behmod	7	8	6	7
	6	7.5	6	

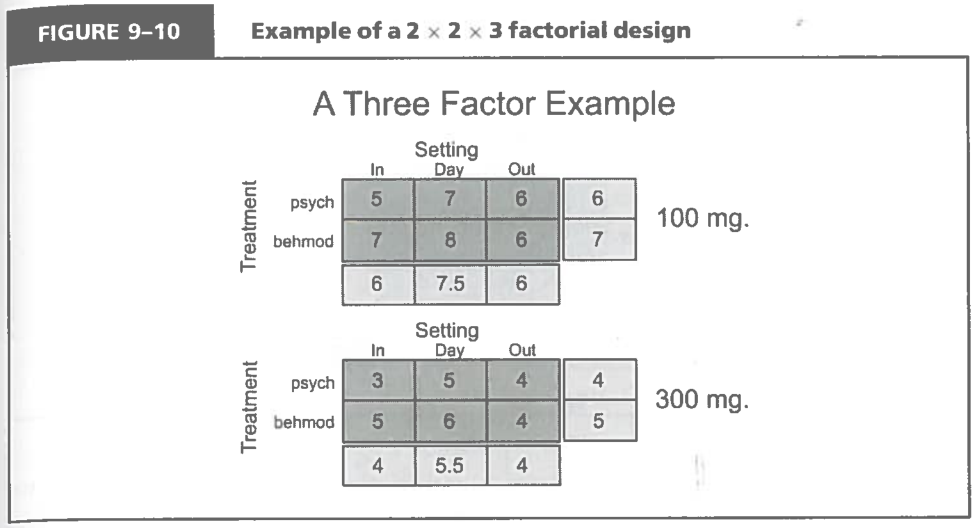
Figure 9–9a shows what an effect for the setting outcome might look like. You have to be careful when interpreting these results because higher scores mean the patient is doing *worse.* It's clear that inpatient treatment works best, day treatment is next best, and outpatient treatment is worst of the three. It's also clear that there is no difference between the two treatment levels (psychotherapy and behavior modification). Even though both graphs in the figure depict the exact same data, it's easier to see the main effect for setting in the graph on the lower left where setting is depicted with different lines on the graph rather than at different points along the horizontal axis.

Figure 9–9b shows a main effect for treatment with psychotherapy performing better (remember the direction of the outcome variable) in all settings than behavior modification. The effect is clearer in the graph on the lower right where treatment levels are used for the lines. Note that in both this and Figure 9–9a, the lines in all graphs are parallel, indicating that there are no interaction effects.

Figure 9–9c shows one possible interaction effect; day treatment is never the best condition. Furthermore, you see that psychotherapy works best with inpatient care, and behavior modification works best with outpatient care.

The other interaction effect shown in Figure 9–9d is a bit more complicated. Although there may be some main effects mixed in with the interaction, what's important here is that there is a unique combination of levels of factors that stands out as superior: psychotherapy done in the inpatient setting. After you identify a best combination like this, the main effects are virtually irrelevant.

FIGURE 9–10 Example of a 2 × 2 × 3 factorial design


A Three Factor Example
Setting
Treatment
100 mg.
300 mg.

Treatment	In	Day	Out		
psych	5	7	6	6	100 mg.
behmod	7	8	6	7	
	6	7.5	6		

Treatment	In	Day	Out		
psych	3	5	4	4	300 mg.
behmod	5	6	4	5	
	4	5.5	4		

A Three-Factor Example Now let's examine what a three-factor study might look like. I'll use the same factors as in the previous example for the first two factors, but here I'll include a new factor for dosage that has two levels. The factor structure in this 2 × 2 × 3 factorial experiment is as follows:

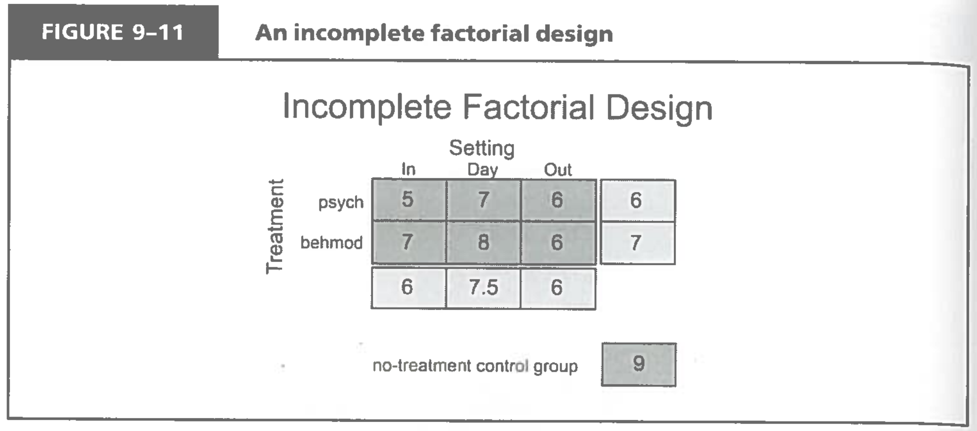
- Factor 1: Dosage
  - 100 mg
  - 300 mg
- Factor 2: Treatment
  - Psychotherapy
  - Behavior modification
- Factor 3: Setting
  - Inpatient
  - Day treatment
  - Outpatient

Notice that in this design you have 2 × 2 × 3 = 12 groups (see Figure 9–10). Although it's tempting in factorial studies to add more factors, the number of groups always increases multiplicatively (is that a real word?). Notice also that to show the tables of means, you have to have two tables that each show a two-factor relationship. It's also difficult to graph the results in a study like this because there will be many different possible graphs. In the statistical analysis, you can look at the main effects for each of your three factors, the three two-way interactions (for example, treatment versus dosage, treatment versus setting, and setting versus dosage) and at the one three-way interaction. Whatever else may be happening, it is clear that one combination of three levels works best: 300 mg and psychotherapy in an inpatient setting. Thus, this study has a three-way interaction. If you were an administrator having to make a choice among the different treatment combinations, you would be best advised to select that one (assuming your patients and setting are comparable to the ones in this study).

Incomplete Factorial Design It's clear that factorial designs can become cumbersome and have too many groups, even with only a few factors. In much research, you won't be interested in a **fully crossed factorial design** like the ones shown previously that pair every combination of levels of factors. Some of the combinations may not make sense from a policy or administrative perspective, or you simply may

fully crossed factorial design A design that includes the pairing of every combination of factor levels.

FIGURE 9-11 An incomplete factorial design



not have the funds to implement all combinations. In this case, you may decide to implement an **incomplete factorial design**. In this variation, some of the cells are intentionally left empty; you don't assign people to get those combinations of factors.

incomplete factorial design
A design in which some cells or combinations in a fully crossed factorial design are intentionally left empty.

One of the most common uses of incomplete factorial design is to allow for a control or placebo group that receives no treatment. In this case, it is actually impossible to implement a group that simultaneously has several levels of treatment factors and receives no treatment at all. So, you consider the control group to be its own cell in an incomplete factorial rubric, which allows you to conduct both relative and absolute treatment comparisons within a single study and to get a fairly precise look at different treatment combinations (see Figure 9–11).

9-4 Randomized Block Designs

randomized block design
Experimental designs in which the sample is grouped into relatively homogeneous subgroups or blocks within which your experiment is replicated. This procedure reduces noise or variance in the data.

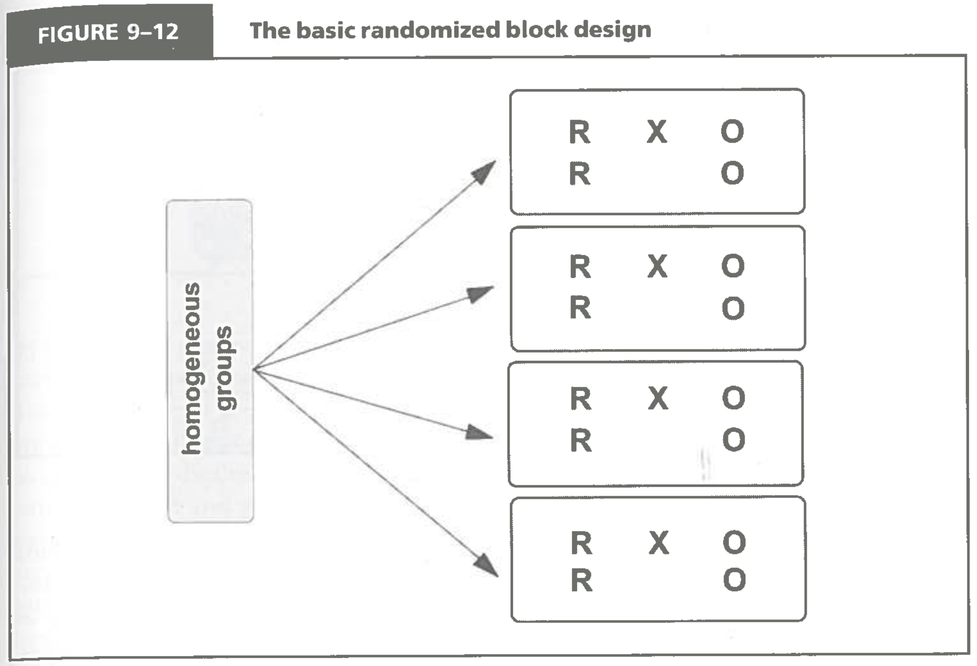
stratified random sampling
A method of sampling that involves dividing your population into homogeneous subgroups and then taking a simple random sample in each subgroup.

The **randomized block design** is research design's equivalent to **stratified random sampling** (see Chapter 2). Like stratified sampling, randomized block designs are constructed to reduce noise or variance in the data (see Section 9-2, Classifying Experimental Designs). How do they do it? They require you to divide the sample into relatively homogeneous subgroups or blocks (analogous to strata in stratified sampling). Then, the experimental design you want to apply is implemented within each block or homogeneous subgroup. The key idea is that the variability within each block is less than the variability of the entire sample. Thus each estimate of the treatment effect within a block is more efficient than estimates across the entire sample. When you pool these more efficient estimates across blocks, you should get a more efficient estimate overall than you would without blocking.

Figure 9–12 shows a simple example. Let's assume that you originally intended to conduct a simple posttest-only, randomized experimental design but you recognized that your sample has several intact or homogeneous subgroups. For instance, in a study of college students, you might expect that students are relatively homogeneous with respect to class or year. So, you decide to block the sample into four groups: freshman, sophomore, junior, and senior. If your hunch is correct—that the variability within class is less than the variability for the entire sample—you will probably get more powerful estimates of the treatment effect within each block. Within each of your four blocks, you would implement the simple post-only randomized experiment.

Notice a couple of things about this strategy. First, to an external observer, it may not be apparent that you are blocking. You implement the same design in each block, and there is no reason that the people in different blocks need to be segregated or separated physically from each other. In other words, blocking doesn't necessarily affect anything that you do with the research participants. Instead, blocking is a strategy for grouping people in your data analysis to reduce noise; it is an *analysis* strategy. Second, you will only benefit from a blocking design if you are correct in your hunch that the blocks are more homogeneous than the entire sample is. If you are wrong—if different college-level classes aren't relatively homogeneous with respect to your measures—you will actually be hurt by blocking. (You'll get a less powerful estimate of the treatment effect.) How do you know whether blocking is a good idea? You need to consider carefully whether the groups are relatively homogeneous. If you are measuring political attitudes, for instance, is it reasonable to believe that freshmen are more like each other than they are like sophomores or juniors? Would they be more homogeneous with respect to measures related to drug abuse? Ultimately the decision to block involves judgment on the part of the researcher.

FIGURE 9-12 The basic randomized block design



9-4a How Blocking Reduces Noise

So how does blocking work to reduce noise in the data? To see how it works, you have to begin by thinking about the nonblocked study. Figure 9–13a shows the pretest-posttest distribution for a hypothetical pre-post randomized experimental design. The plus symbol indicates a program group case, and the circle symbol signifies comparison-group members. You can see that for any specific pretest value, the program group tends to outscore the comparison group by about 10 points on the posttest (meaning, there is about a 10-point posttest mean difference).

Now, let's consider an example that divides the sample into three relatively homogeneous blocks. To see what happens graphically, you'll use the pretest measure to block. This ensures that the groups are homogeneous. Let's look at what is happening within the third block (see Figure 9–13b). Notice that the mean difference is still the same as it was for the entire sample—about 10 points within each block. Also notice that the variability of the posttest is much less within the block than it is for the entire sample. Remember that the treatment effect estimate is a signal-to-noise ratio. The signal in this case is the mean difference. The noise is the variability. Figure 9–13a and Figure 9–13b show that you haven't changed the signal by moving to blocking; there is still about a 10-point posttest difference. However, you have changed the noise; the variability on the posttest is much smaller within each block than it is for the entire sample. So, the treatment effect will have less noise for the same signal.

FIGURE 9–13a Pre-post distribution for a randomized experimental design without blocking

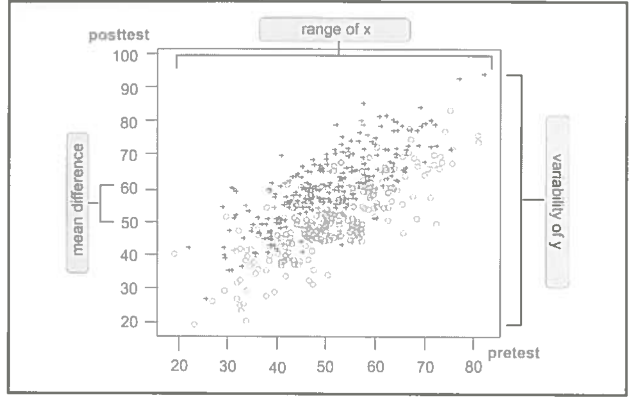
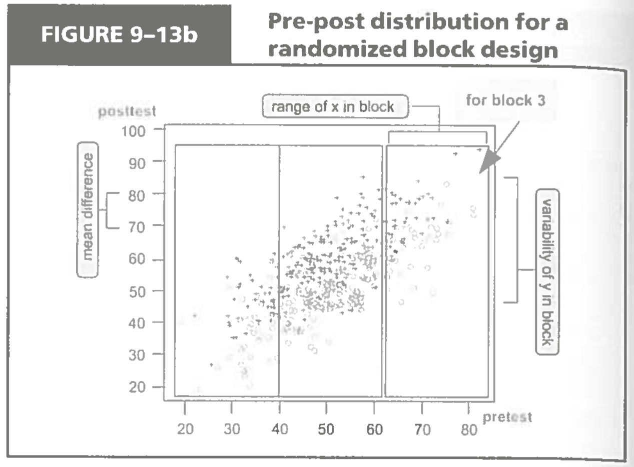


FIGURE 9–13b Pre-post distribution for a randomized block design



It should be clear from the graphs that the blocking design in this case yields the stronger treatment effect. However, this is true only because the blocks were homogeneous. If the blocks weren't homogeneous—their variability was as large as the entire sample's—you would actually get worse estimates than in the simple randomized experimental case. You'll see how to analyze data from a randomized block design in Section 14-3c, Randomized Block Analysis.

9-5 Covariance Designs

The basic analysis of covariance design (ANCOVA or ANACOVA) is a pretest-posttest randomized experimental design. The notation shown in Figure 9–14 suggests that the pre-program measure is the same one as the post-program measure (otherwise, you would use subscripts to distinguish the two), so you would call this a pretest. Note however that the pre-program measure doesn't have to be a pretest; it can be any variable measured prior to the program intervention. It is also possible for a study to have more than one **covariate**.

covariate
Variables you adjust for in your study.

The pre-program measure or pretest is sometimes also called a covariate because of the way it's used in the data analysis; you co-vary it with the outcome variable or posttest to remove variability or noise. Thus, the ANCOVA design falls in the class of a noise-reduction experimental design (see Section 9-2, Classifying Experimental Designs).

In social research, you frequently hear about statistical adjustments that attempt to control for important factors in your study. For instance, you might read that an analysis examined posttest performance after *adjusting for* the income and educational level of the participants. In this case, *income* and *education level* are covariates. Covariates are the variables you *adjust for* in your study. Sometimes the language that will be used is that of *removing the effects* of one variable from another. For instance, you might read that an analysis examined posttest performance after *removing the effect of* income and educational level of the participants.

9-5a How Does a Covariate Reduce Noise?

One of the most important ideas in social research is how you make a statistical adjustment—adjust one variable based on its covariance with another variable. If you understand this idea, you'll be well on your way to mastering social research. What I want to do here is show you a series of graphs that illustrate pictorially what it means to adjust for a covariate.

FIGURE 9–14 Notation for the basic analysis of covariance design

R O X O
R O O

Let's begin with data from a simple ANCOVA design as described previously. Figure 9–15a shows the pre-post bivariate distribution. Each dot on the graph represents the pretest and posttest score for an individual. The plus signifies a program or treated case, and a circle describes a control or comparison case. You should be able to see a few things immediately. First, you should be able to see a whopping treatment effect! It's so obvious that you don't even need statistical analysis to tell you whether there's an effect (although you may want to use statistics to estimate its size and probability). How do I know there's an effect? Look at any pretest value (value on the horizontal axis). Now, look up from that value; you are looking up the posttest scale from lower to higher posttest scores. Do you see any pattern with respect to the groups? It should be obvious to you that the program cases (the plusses) tend to score higher on the posttest at any given pretest value. Second, you should see that the posttest variability has a range of about 70 points.

Figure 9–15b shows the graph with straight lines fitted to the data. The lines on the graph are regression lines that describe the pre-post relationship for each of the groups. The regression line shows the expected posttest score for any pretest score. The treatment effect is even clearer with the regression lines. You should see that the line for the treated group is about 10 points higher than the line for the comparison group at any pretest value.

What you want to do is remove some of the variability in the posttest while preserving the difference between the groups. In other terms, you want to *adjust* the posttest scores for pretest variability. In effect, you want to *subtract out* the pretest. You might think of this as subtracting the line from each group from the data for each group. How do you do that? Well, why don't you actually subtract? Find the posttest difference between the line for a group and each actual value (Figure 9–15c). Each of these differences is called a *residual;* it's what's left over when you subtract a line from the data.

Now, here comes the tricky part. What does the data look like when you subtract out a line? You might think of it almost like turning the graph in Figure 9–15c clockwise until the regression lines are horizontal. Figure 9–15d and Figure 9–15e show this in two steps. First, you construct an *x-y* axis system, where the *x* dimension is parallel to the regression lines.

Then, you actually turn the graph clockwise so that the regression lines are flat horizontally (Figure 9–15e). Notice how big the posttest variability or range is in Figure 9–15e (as indicated by the double arrow). You should see that the range is considerably smaller than the 70 points with which you started. You should also see that the difference between the lines is the same as it was before. So, you have in effect reduced posttest variability while maintaining the group difference. You've lowered the noise while keeping the signal at its original strength. The statistical adjustment procedure will result in a more efficient and more powerful estimate of the treatment effect.

You should also note the shape of the pre-post relationship. Essentially, the plot now looks like a zero correlation between the pretest and posttest and, in fact, it is. How do I know it's a zero correlation? Because any line that can fit through the data well would be horizontal. There's no slope or relationship, and there shouldn't be. This graph shows the pre-post relationship after you've removed the pretest! If you've removed the pretest from the posttest there will be no pre-post correlation left.

FIGURE 9–15a A pre-post distribution for a covariance design

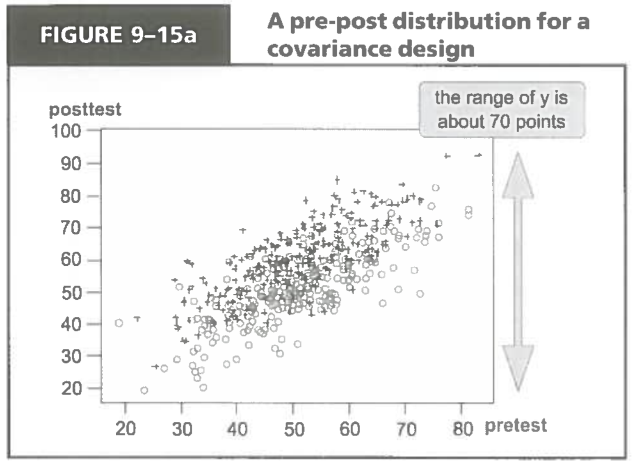

the range of y is about 70 points

FIGURE 9–15b Pre-post distribution for a covariance design with regression lines fitted

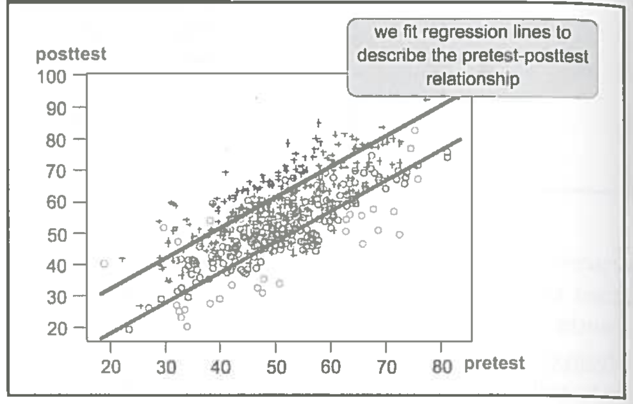

we fit regression lines to describe the pretest-posttest relationship

FIGURE 9–15c Subtract the posttest value from the predicted posttest value to obtain the residual for a single participant (point)

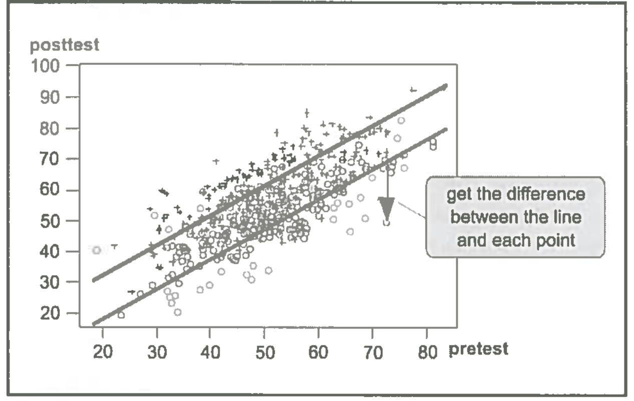
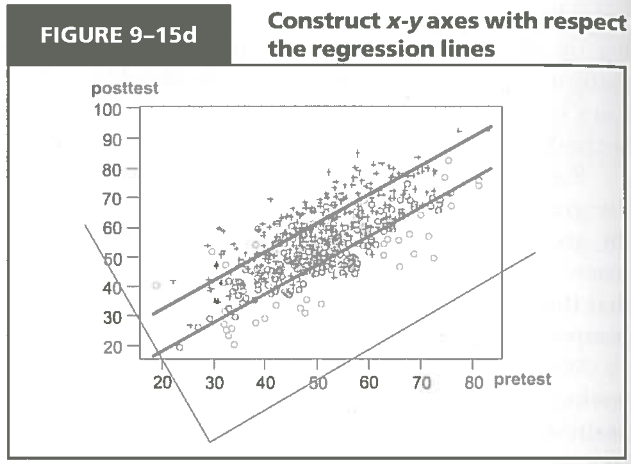

get the difference between the line and each point

FIGURE 9–15d Construct *x-y* axes with respect to the regression lines



Finally, redraw the axes to indicate that the pretest

FIGURE 9–15e The rotated view of Figure 9-15d with the pre-post relationship removed

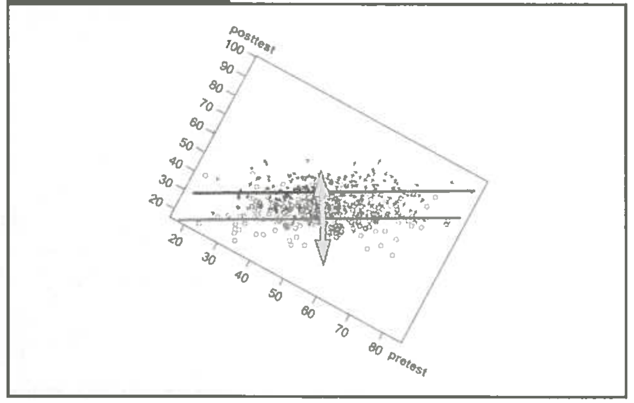
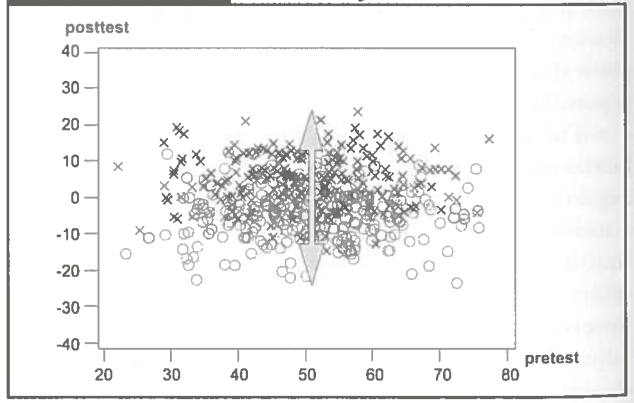


FIGURE 9–15f The pre-post distribution when the pre-post relationship has been adjusted out



has been removed. Figure 9–15f shows the posttest values as the original posttest values minus the line (the predicted posttest values). That's why you see that the new posttest axis has 0 at its center. Negative values on the posttest indicate that the original point fell below the regression line on the original axis. Here, the posttest range is about 50 points instead of the original 70, even though the difference

between the regression lines is the same. You've lowered the noise while retaining the signal.

Disclaimer: Okay, I know some statistical hotshot out there is fuming about the inaccuracy in my previous description. My picture rotation is not exactly what you do when you adjust for a *covariate*. My description suggests that you drop perpendicular lines from the regression line to each point to obtain the subtracted difference. In fact, you drop lines that are perpendicular to the horizontal axis, not the regression line itself (in least squares regression you are minimizing the sum of squares of the residuals on the dependent variable, not jointly on the independent and dependent variable as suggested here). In any event, although my explanation may not be perfectly accurate from a statistical point of view, it's not far off, and it conveys clearly the idea of subtracting out a relationship. I thought I'd just put this disclaimer in to let you know I'm not dumb enough to believe that my description is perfectly accurate.

The adjustment for a covariate in the ANCOVA design is accomplished with the statistical analysis, not through rotation of graphs. (See Section 14-3d, Analysis of Covariance, for details).

9-5b Summary

Here are some thoughts to conclude this topic. The ANCOVA design is a noise-reducing experimental design. It *adjusts* posttest scores for variability on the covariate (pretest); this is what it means to *adjust* for the effects of one variable on another in social research. You can use *any* continuous variable as a covariate, but the pretest is usually best. Why? Because the pretest is usually the variable that is most highly correlated with the posttest. (A variable should correlate highly with itself, shouldn't it?) Because it's so highly correlated, when you subtract it out or remove it, you're removing extraneous variability from the posttest. The rule in selecting covariates is to select the measure(s) that correlate most highly with the outcome and, for multiple covariates, have little intercorrelation. (Otherwise, you're simply adding redundant covariates and you actually lose precision by doing that.) For example, you probably wouldn't want to use both gross and net income as two covariates in the same analysis because they are highly related and therefore redundant as adjustment variables.

9-6 Hybrid Experimental Designs

Hybrid experimental designs are just what the name implies—new strains that are formed by combining features of more established designs. Many variations can be constructed from standard design features. Here, I'm going to introduce two hybrid designs. I'm featuring these because they illustrate especially well how a design can be constructed to address specific threats to internal validity.

9-6a The Solomon Four-Group Design

The **Solomon four-group design** is designed to deal with a potential **testing threat to internal validity** discussed in Chapter 7. Recall that a testing threat occurs when the act of taking a test affects how people score on a retest or posttest. The design notation is shown in Figure 9–16. It's probably not a big surprise that this design has four groups. Note that two of the groups receive the treatment and two do not. Furthermore, two of the groups receive a pretest and two do not. One way to view this is as a 2 × 2 (Treatment Group *X* Measurement Group) factorial design. Within each treatment condition, one group is pretested and one is not. By explicitly including testing as a factor in the design, you can assess experimentally whether a testing threat is operating.

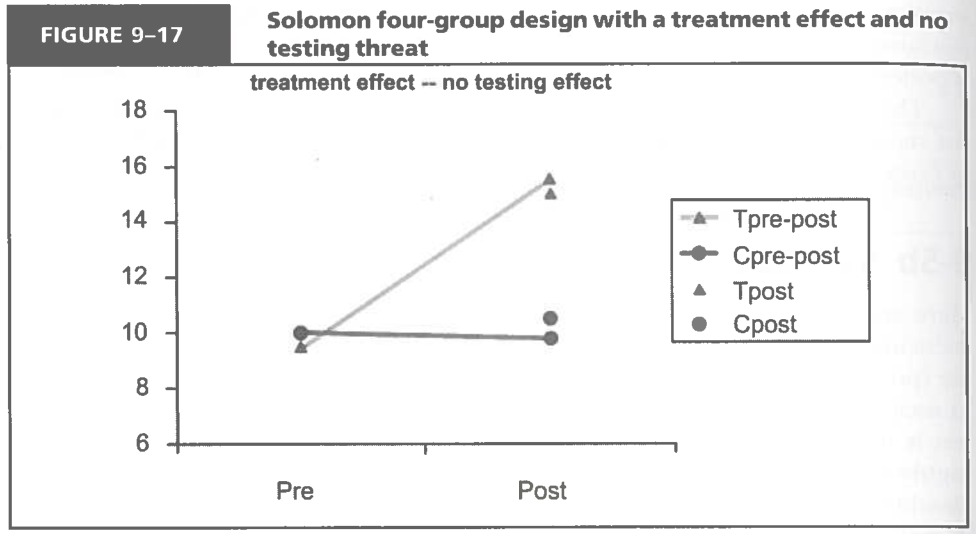
Solomon four-group design
This design has four groups. Two of the groups receive the treatment and two do not. Furthermore, one of the treatment groups and one of the controls receive a pretest and the other two do not. By explicitly including testing as a factor in the design, you can assess experimentally whether a testing threat is operating.

testing threat to internal validity
A threat to internal validity that occurs when taking the pretest affects how participants do on the posttest.

FIGURE 9–16 Design notation for the Solomon four-group design

R	O	X	O
R	O		O
R		X	O
R			O

FIGURE 9–17 Solomon four-group design with a treatment effect and no testing threat



Let's look at a couple of possible outcomes from this design. The first outcome graph (Figure 9–17) shows what the data might look like if there is a treatment or program effect and no testing threat. You need to be careful when interpreting this graph to note that there are six dots: one to represent the average for each *O* in the design notation. To help you visually see the connection between the pretest and posttest average for the same group, a line connects the dots. The two dots that are not connected by a line represent the two post-only groups. Look first at the two pretest means. They are close to each other because the groups were randomly assigned. On the posttest, both treatment groups outscored both controls. Now, look at the posttest values. There appears to be no difference between the treatment groups, even though one got a pretest and the other did not. Similarly, the two control groups scored about the same on the posttest. Thus, the pretest did not appear to affect the outcome. However, both treatment groups clearly outscored both controls. There is a main effect for the treatment.

Now, look at a result with evidence of a testing threat (Figure 9–18). In this outcome, the pretests are again equivalent (because the groups were randomly assigned). Each treatment group outscored its comparable control group. The pre-post treatment outscored the pre-post control, and the post-only treatment outscored the post-only control. These results indicate that there is a treatment effect. However, here both groups that had the pretest outscored their comparable non-pretest group. That's evidence of a testing threat.

switching-replications design
A two-group design in two phases defined by three waves of measurement. The implementation of the treatment is repeated in both phases. In the repetition of the treatment, the two groups switch roles: The original control group in phase 1 becomes the treatment group in phase 2, whereas the original treatment acts as the control. By the end of the study, all participants have received the treatment.

9-6b Switching-Replications Design

FIGURE 9–18 Solomon four-group design with both a treatment effect and a testing threat

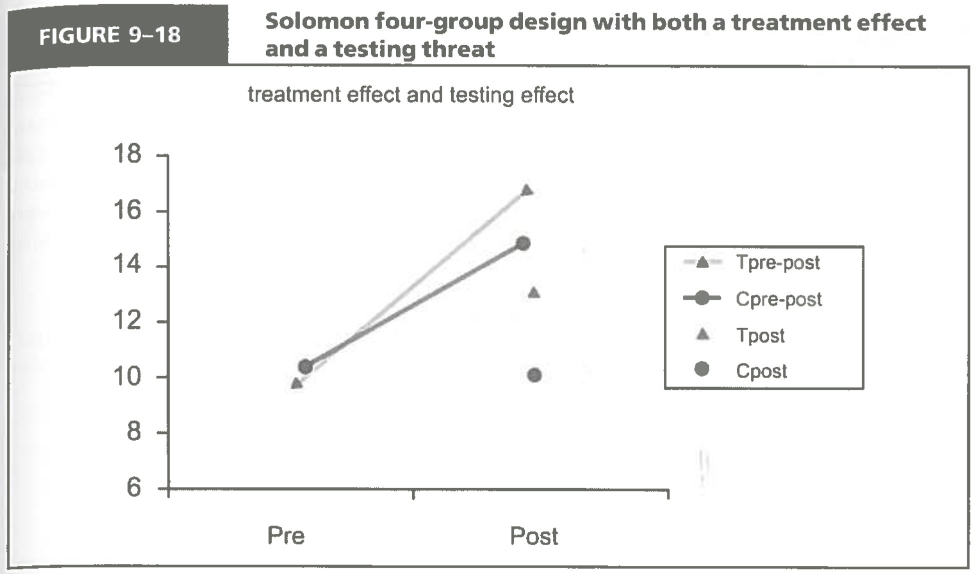


FIGURE 9–19 Notation for the switching-replications, randomized experimental design

R	O	X	O		O
R	O		O	X	O

The **switching replications design** is one of the strongest of the experimental designs. When the circumstances are right for this design, it addresses one of the major problems in experimental designs: the need to deny the program to some participants through random assignment. The design notation (see Figure 9–19) indicates that this is a two-group design with three waves of measurement. You might think of this as two pre-post, treatment-control designs grafted together. That is, the implementation of the treatment is repeated or *replicated*. In the repetition of the treatment, the two groups *switch* roles; the original control group becomes the treatment group in phase 2, whereas the original treatment acts as the control. By the end of the study, all participants have received the treatment.

The switching-replications design is most feasible in organizational contexts where programs are repeated at regular intervals. For instance, it works especially well in schools that are on a semester system. All students are pretested at the beginning of the school year. During the first semester, Group 1 receives the treatment, and during the second semester, Group 2 gets it. The design also enhances organizational efficiency in resource allocation. Schools need to allocate only enough resources to give the program to half of the students at a time.

Let's look at two possible outcomes. In the first example, the program is given to the first group, and the recipients do better than the controls (see Figure 9–20). In the second phase, when the program is given to the original controls, they catch up to the original program group. Thus, you have a converge-diverge-reconverge outcome pattern. You might expect a result like this when the program covers specific content that the students master in the short term and where you don't expect them to continue improving as a result.

Now, look at the other example result (see Figure 9–21). During the first phase, you see the same result as before; the program group improves while the control group does not. As before, during the second phase, the original control group, in this case the program group, improved as much as the first program group did. This time, however, during phase two the original program group continued to increase even after it no longer received the program. Why would this happen? It could happen in circumstances where the program has continuing effects. For instance, if the program focused on learning skills, students might continue to

FIGURE 9–20 **Switching-replications design with a short-term persistent treatment effect**

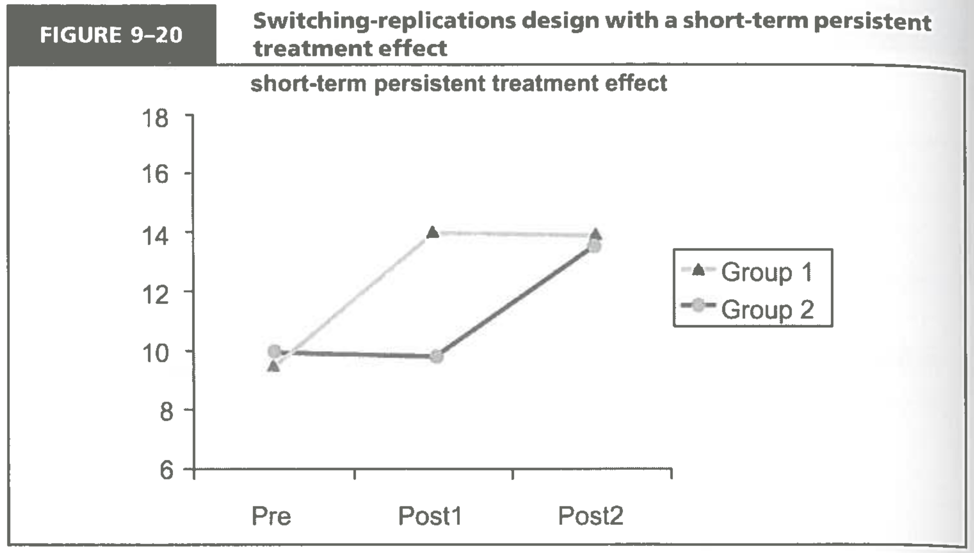
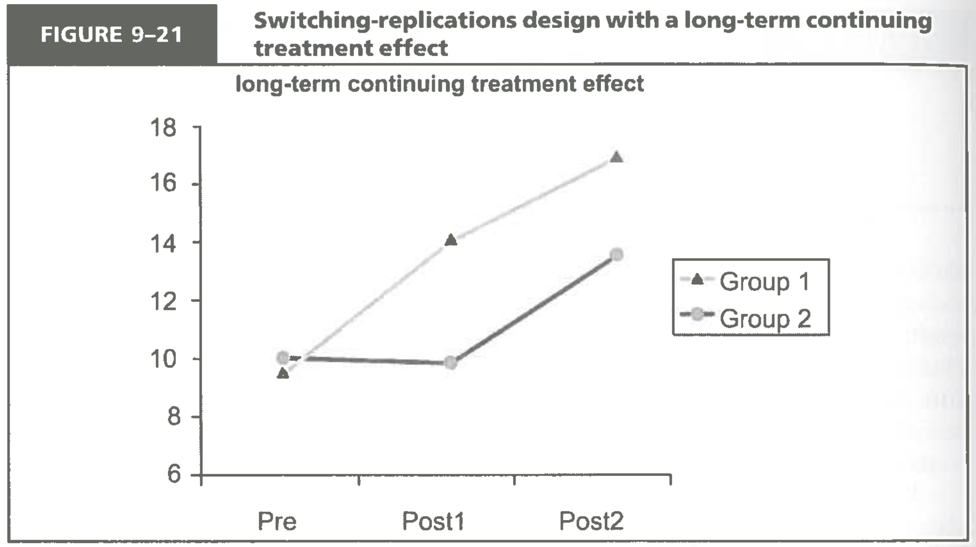


FIGURE 9–21 **Switching-replications design with a long-term continuing treatment effect**



improve even after the formal program period because they continue to apply the skills and improve in them.

I said earlier that both the Solomon four-group and the switching-replications designs addressed specific threats to internal validity. It's obvious that the Solomon design addressed a testing threat. But what does the switching-replications design address? Remember that in randomized experiments, especially when the groups are aware of each other, there is the potential for social threats to internal validity; compensatory rivalry, compensatory equalization, and resentful demoralization are all likely to be present in educational contexts where programs are given to some students and not to others. The switching-replications design helps mitigate these threats because it ensures that everyone will eventually get the program. In addition, it allocates who gets the program first in the fairest possible manner, through the lottery of random assignment.

Summary

This chapter introduced experimental designs. The basic idea of a randomized experiment was presented along with consideration of how it addresses internal validity, the key concepts

of probabilistic equivalence, and the distinction between random selection and random assignment. Experimental designs can be classified as signal enhancers or noise reducers. Factorial designs were presented as signal enhancers that emphasize studying different combinations of treatment (signal) features. Two types of noise-reducing strategies—randomized blocks and covariance designs—were presented along with descriptions of how each acts to reduce noise in the data. Finally, two hybrid experimental designs—the Solomon four-groups and switching-replications designs—were presented to illustrate the versatility of experimental designs and the ability to tailor designs that address specific threats to internal validity.

Login to the Online Edition of your text at www.atomicdog.com to find additional resources located in the Study Guide at the end of each chapter.