SAMPLE CHAPTER

EDUCATIONAL RESEARCH IN AN AGE OF ACCOUNTABILITY

Robert E. Slavin

ISBN 0-205-43982-9

SAMPLE CHAPTER 2: RANDOMIZED EXPERIMENTAL DESIGNS

The pages of this Sample Chapter may have slight variations in final published form.

PEARSON
PEARSON Longman
ALLYN & BACON/LONGMAN
www.ablongman.com

2 Randomized Experimental Designs

experimental comparison design: An experimental design that allows for the comparison of one treatment condition with another on two or more different groups.

Does ability grouping produce better achievement than heterogeneous assignment to classes? Does cooperative learning improve student self-esteem more than traditional methods? Does rewarding students for reading increase or decrease their interest in reading? These and other questions that involve comparisons of one treatment condition with another are usually best answered by **experimental comparison designs.** These are studies in which subjects are assigned by the experimenter (usually randomly) to two or more groups, different treatments are applied to the different groups, and the effects of the treatments on one or more outcomes (dependent variables) are measured. This chapter discusses the logic of experimental comparisons and the practical problems of doing such studies in real schools and classrooms.

Experimental Comparisons in an Age of Accountability

Shortly after the American Revolution, wheat farmers in the Piedmont region of Virginia and Maryland began to notice that each year, their wheat yields were diminishing. Things were getting so bad that many were abandoning their farms and moving west. One Virginia farmer, John Binns, had heard about remarkable results from plowing crushed limestone into wheat fields. He bought some limestone and spread it on some of his fields; then he planted some wheat on fields with limestone and some wheat on fields without limestone. The results were astonishing: The limestone-treated fields yielded much more wheat. Later, Binns found that if he planted clover on limestone-treated fields and grazed cattle on the clover, his wheat yields the following year were even greater. Over time, he developed a crop rotation method that made the Piedmont the breadbasket of the emerging United States.

Binns's method for testing the effects of limestone on wheat yields is an example of an experimental comparison, one of the most important experimental designs used in educational research. Binns randomly chose some of his fields to be treated and some to remain as they had been before. Today, we would call the treated fields the **experimental group,** because they were receiving the experimental treatment, and the untreated fields the **control group.** Choosing fields at random to be treated or not treated was important; otherwise, Binns might have inadvertently chosen a field that had better soil anyway, even without limestone, for one or the other group.

Binns's experimental design had all the elements that are important in educational experiments. He made sure the treated and untreated fields were the same on all factors except their limestone content, and he carefully measured the output of each type of field. In his experiments, he could be certain that the differences he saw in wheat yields were due to limestone, not to other factors.

Educational researchers can conduct experiments in exactly the same way. They can randomly assign students or classes to receive a treatment (the experimental group) and others to continue learning as they had before (the control group) and then carefully measure the outcomes on tests or other assessments. However, people are a lot more complex than wheat, and schools have many characteristics that sometimes make simple experiments difficult or impossible.

experimental group: A group assigned to receive some experimental treatment.

control group: A group assigned to be untreated or to receive a treatment other than the experimental treatment.

Randomized experimental comparison designs have taken on great importance in recent years in the evaluation of educational programs. The Institute of Education Sciences (IES), the research arm of the U.S. Department of Education, is undertaking a campaign to greatly increase the use of randomized experiments, and its What Works Clearinghouse gives a strong preference to programs that have been successfully evaluated using such designs. Randomized experiments are referred to as meeting the "gold standard" for research design. No one denies that other designs are valuable for answering a variety of important research questions, but when it comes to "what works" questions, such as evaluations of particular teaching programs and practices, IES strongly recommends randomized experiments (see U.S. Department of Education, 2004).

Random Assignment

experimental treatment: A treatment applied to some subjects in an experimental comparison design whose effects on one or more dependent (outcome) variable or variables are to be contrasted with the effects of other treatments or control (untreated) conditions.

selection bias: Any nonrandom factor that might influence the selection of individuals into one or another treatment.

selection effects: Effects on outcomes of preexisting differences between subjects in experimental and control groups.

random assignment: Selection into one or another treatment (or control) group in an experimental comparison design by chance, in such a way that all individuals to be assigned have a known and equal probability of being assigned to any given group.

One of the most important features of most experimental comparison designs is the use of random assignment of subjects to the various treatments. Some subjects (students, for example) are assigned to receive one **experimental treatment,** while others are assigned to receive a different treatment. Some students might be assigned to study mathematics using cooperative learning, while other students are assigned to receive mathematics instruction in the form of lectures. If students are randomly assigned to treatments, the experimenter determines which students will be in which treatments by a chance process. For example, flipping a coin could determine whether a student goes into the cooperative learning class or into the lecture class within each pair of students. All those who get heads might be assigned to the lecture class.

Random assignment solves one of the most critical problems of research design: **selection bias,** which is a nonrandom factor that might influence the selection of individuals into one or another treatment. One of the biggest problems in learning from studies that do not use random assignment is the difficulty in separating **selection effects** (effects on outcomes of preexisting differences between subjects in experimental and control groups) from treatment effects. Does Jones High School really do the best job of teaching in the city, or does it simply have the best students? Do good coaches make good teams, or do good players make their coaches look good by winning? Does small class size increase student achievement, or do more able students tend to find themselves more often in small classes? Whenever we wish to compare the effect of one treatment to another, we must be sure that the subjects in each treatment are reasonably equal (on the average) on all important criteria. Otherwise, unequal selection effects, or selection bias, may make any differences we find between treatments uninterpretable.

Random assignment to different treatment conditions virtually rules out selection bias as an explanation for differences between treatments, making it one of the best ways to avoid getting false positive and false negative errors. The essence of random assignment is that there is no way to tell in advance who will receive each treatment. Selection into one or another treatment (or control) group is done by chance and in such a way that all individuals to be assigned have a known and equal probability of being assigned to any given group. For example, a researcher might take a list of 100 children, put their names on slips of paper in a box, mix up

"I think you'd do better in my research methods class if you did fewer random assignments."

the slips, and draw names at random, putting half of the slips in one pile and half in another. The children whose names are in the first pile will see a film designed to improve their attitudes toward Mexican Americans; the other students will see a film unrelated to Mexican Americans. At the end of the study, all students will complete a questionnaire on attitudes toward Mexican Americans.

What is important about random assignment in this study is that picking the students at random for the two conditions answers most questions about the equivalence of the two groups before they saw the movies. If the group that saw the film on Mexican Americans does have more positive attitudes toward Mexican Americans, as measured on the attitude scale given after the film, no one (not even the Gremlin) could argue that this happened because the selected students were brighter, more tolerant, more experienced with Mexican Americans, or from more liberal families than the other students. All students had an equal chance to be chosen for either group, and the number of students in each condition (50) is large enough to make it almost certain that the groups will be very close to equal on these and other factors.

RANDOM ASSIGNMENT OF INDIVIDUALS

Random assignment can be done in many ways. We might use a table of random numbers (one appears in Appendix 5). To do this, we might list all students in alphabetic order and then decide by flipping a coin that even numbers on the table of random numbers represent the experimental group and that odd numbers represent the control group. We would then choose a random starting place in the random numbers table. Let's say the random numbers table began as follows:

7608213143295835 . . .

To do the random assignment, we would assign the first student to the control group (because 7 is odd), the next student to the experimental group (6 is even), the next to the experimental group (0 is even), and so on. We would continue this process until all 100 students had been assigned.

STRATIFIED RANDOM ASSIGNMENT

Although random assignment usually produces groups that can be considered equal, there is no guarantee that the groups will in fact be equal on every relevant factor. By chance, it is possible that the groups will be different in some important way, especially if the number of subjects in each group is less than 30 or so. Whenever it is possible to obtain data on each subject on variables that could be related to the outcomes we are studying, particularly when the number of students in each group is small, random assignment should be stratified on these variables.

Stratified random assignment means that students are randomly assigned within a particular category, or stratum. In our example regarding the Mexican American film, we might want to make sure that there were equal numbers of boys and girls in each treatment group, because we suspect that boys and girls might have different attitudes or might be affected differently by the film. Let's assume that there were 56 boys and 44 girls. We might have randomly assigned the boys (28 boys to each treatment group) and then the girls (22 to each group), guaranteeing that the two groups would have equal numbers of boys and girls. If there were African Americans and whites in the sample, we might have randomly assigned students within the subsamples of African American males, African American females, white males, and white females, thereby stratifying on two variables: sex and race.

In research on student achievement, the most important variable we need to be sure is equal in different treatment groups is prior academic achievement level. Because students' learning rates depend to a large degree on how much they have learned in the past, even small group differences on prior achievement tests (or similar measures) can make meaningful interpretation of differences in **posttest** measures difficult. (A posttest is a test or questionnaire given at the end of some treatment period.)

Figure 2.1 on page 30 shows how a class of 31 students might be randomly assigned to two treatment groups, stratifying on academic achievement level and sex. The boys and girls were separately ranked, based on their most recent test scores. Then they were placed in matched pairs for random assignment (for example, Sam and Tyrone are the highest-scoring boys, Paula and Laura are the lowest-scoring girls). To assign the students to the experimental (checked) or control (unchecked) group, a coin was flipped for each pair. If the coin came up heads, the top student in the pair was assigned to the experimental group; if tails, the bottom student. Note that one student, Maria, was left over in the matching. We made sure that the leftover student was average in past achievement, so that it would make no difference to which group she was assigned. We flipped our coin and assigned her to the experimental group.

stratified random assignment: Random assignment of subjects to one or more groups done in such a way as to ensure that each group will have certain characteristics.

posttest: A test or questionnaire given at the end of some treatment period.

Stratifying on sex and achievement makes it certain that the experimental and control groups will have very nearly equal numbers of boys and girls and of high and low achievers, but the groups are still randomly assigned because there was no way to predict who would be in each group. Not using random assignment increases the chance that we will make either a false positive or a false negative error. For example, let's say that instead of randomly assigning students to treatments, we had shown the attitude-improvement film to Ms. Jackson's first-period class and the neutral film to her fifth-period class. If differences favoring the students who saw the attitude-change film are found, they could well be due to the fact that the

FIGURE 2.1

Example of Random Assignment Stratifying on Achievement Level and Sex

(Highest) Achievement Rank	Boys	Girls
	1. Sam	1. Amanda ✓
	2. Tyrone ✓	2. Sylvia
	3. Raoul ✓	3. Staphanie ✓
	4. Alan	4. Isabel
	5. Todd ✓	5. Gwynn
	6. Ivan	6. Natalie ✓
	7. Frank ✓	7. Maria ✓
	8. Isaac	8. Evelyn
	9. Eric	9. Noreen ✓
	10. David ✓	10. Teresa
	11. Richard	11. Ellen ✓
	12. Antonio ✓	12. Xandra ✓
	13. Nathan ✓	13. Melissa
	14. Dan	14. Paula
	15. Otto	15. Laura ✓
(Lowest)	16. Mack ✓	

Note: A check (✓) indicates that the student was randomly assigned to the experimental group.

students in the first-period class already had more positive attitudes than those in the fifth-period class. This would produce a false positive error. If differences are not found, it could be that the film (which was, let's assume, effective in improving attitudes) had the effect of making the class with poor attitudes resemble the one with good attitudes, so the failure to find statistically significant differences would be a false negative error.

Unfortunately, random assignment is often impossible in social science research. For example, in a study about differences in creativity between boys and girls, it is of course impossible to randomly assign students to *be* boys or girls.

Furthermore, school administrators are usually reluctant to randomly assign students to classes for any substantial period of time, and it is almost impossible to randomly assign students to schools. On the other hand, it is often relatively easy to randomly assign *classes* or *teachers* to different treatments in educational research, and there are ways to deal with the problems of nonrandom assignment. (These are discussed later in this chapter.) However, whenever there is a deviation from true random assignment of many individuals, the burden of proof is on the researcher to demonstrate that the groups being compared can truly be considered equivalent.

Randomized Experimental Comparisons

The ideal experimental design is the true experiment, or randomized experimental comparison, in which individuals are randomly assigned to one or more treatment conditions, the treatments are applied, and the results are compared. A **pretest** may be given at the beginning of the study. This is not absolutely necessary, however, since the random assignment is likely to produce equal groups, especially if the number of subjects in each group is large and/or the randomization was stratified on characteristics that must be equal across treatments (such as achievement level, sex, or race/ethnicity, depending on the topic of the research).

As an example of a randomized experimental comparison, consider the case of a researcher who wants to find out whether students learn better from text if they are allowed to talk about it with a classmate (classmate discussion) than if the teacher conducts a class discussion on it (teacher-led discussion). The researcher locates eight fifth-grade classes in several elementary schools and randomly assigns students to eight new classes, stratifying on reading achievement test scores to be sure that the groups are equal on this critical variable. Four classes are assigned to receive the classmate discussion treatment, and four are assigned to receive the teacher-led discussion treatment. The teachers are also randomly assigned to the two treatments. Each day, the students spend 20 minutes reading short stories and 20 minutes either discussing the stories with partners or participating in a whole-class discussion, depending on the treatment. After four days, all students are tested on their recall of the main ideas from the stories.

In terms of internal validity, this is a good study. The groups can be definitely considered equivalent, because students were randomly assigned to groups stratifying on reading achievement. If the groups turn out to differ significantly on the posttest, we can be relatively confident that the difference in treatments accounts for the difference in outcomes, not any preexisting differences between students or other external factors. This is the beauty of the true experiment: The design of the study rules out most explanations for findings other than that the treatments made the difference. However, although there is no better experimental design than a randomized experimental comparison (all other things being equal), this is not to say that use of such a design is a guarantee of internal validity.

pretest: A test or questionnaire given before some treatment begins.

Consider the previous example, a true experiment. What if the researcher had used only one class per treatment instead of four? In this case, it would have been impossible to separate **teacher effects** (for example, the ability of the teacher to organize the class and present material) and **class effects** (for example, the effects of students on each other) from true treatment effects. (See Chapter 10 for more on teacher and class effects.) This effect, which is called **confounding,** occurs when two variables are so mixed up with each other that it's impossible to tell which is responsible for a given outcome. Since we are interested in the treatments and not in the qualities of individual teachers, teacher effects are pure nuisance. Even with four teachers per treatment, teacher and class effects cannot be ruled out as an explanation for any findings. In fact, quantitative methodologists would insist that if class is the unit that is randomly assigned, then class (and teacher) should be the unit of analysis (requiring 40 to 50 teachers) to avoid the problem with teacher and class effects.

One way to deal with teacher effects in experiments is to rotate teachers across the different treatments. In our example, each teacher could spend some time teaching the classmate discussion group and some time teaching the teacher-led discussion group. This might help reduce the impact of teacher effects, but rotating teachers across classes has its problems, too. Although this procedure reduces the chance that teacher effects will be confounded with treatment effects, it could be argued that the rotation itself becomes part of both treatments and might impact the two treatments differently.

What this discussion is meant to convey is that in field research, it is difficult to guarantee an unassailable design. The classmate discussion study described earlier is in many ways a good study. Because it is an experiment with random assignment, there would be little possibility that preexisting differences between students caused any difference between the groups. The topic is important and builds on prior literature. This does not mean that the criticisms discussed are invalid; it does mean that the study itself is strong enough to be a basis for further research that might answer the criticisms. These criticisms illustrate, however, that while randomized experimental comparisons have important strengths, by themselves, they are in no way a guarantee of adequate internal validity. Only educated common sense can tell us when a study adequately answers the questions it poses.

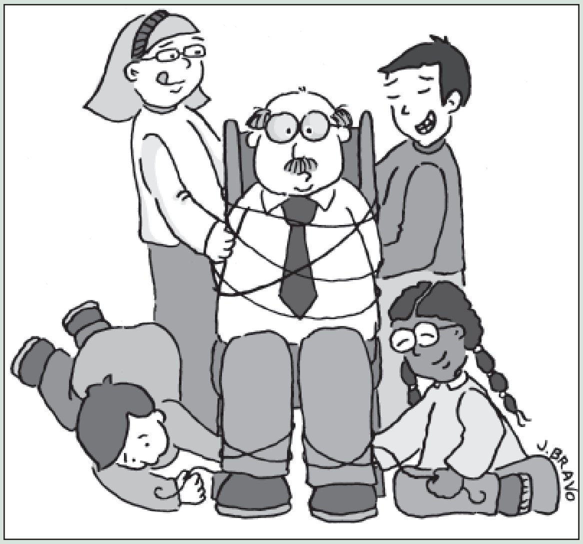
teacher effects: The effects on students of having a particular teacher.

class effects: The effects on students of being in a certain class.

confounding: A situation in which the independent effects of two or more variables cannot be determined because the variables cannot be studied separately.

CONTROL GROUPS

A control group, sometimes called a *counterfactual,* is a group assigned to receive a treatment that serves as a point of comparison for one or more treatment groups. The nature of the control group is very important in experimental comparisons. Sometimes, the control group is intended to represent what students would have experienced if the experiment had not taken place. Research articles often describe such a control group as receiving "traditional instruction" or "standard textbooks." Otherwise, a control group might receive an alternative, widely accepted treatment that might be considered current best practice. For example, a study of a new one-



"I wonder if it's too late to get into the control group?"

to-one tutoring model might provide a program called Reading Recovery to the control group, because Reading Recovery is the most extensively researched and widely used current tutoring model (see Pinnell, DeFord, & Lyons, 1988).

While using traditional instruction as a control group gives researchers a practically meaningful point of comparison, this strategy is often criticized. Teachers who know they are in the control group may feel unmotivated, while those in the treatment group may feel energized. This positive effect of being in an experimental group is called the **Hawthorne effect** (discussed further in Chapter 10). If a well-established and widely used alternative to traditional instruction exists, critics might argue that the comparison of any new program should be to best practice, not to common practice. For these reasons, methodologists often prefer a comparison of a new treatment to another treatment, rather than to traditional instruction. The problem with this strategy is that if the new treatment and the alternative do not differ in outcomes, we don't know if they were both effective or both ineffective. If the new treatment has better results, advocates for the alternative treatment are sure to argue that their favored program was poorly implemented. The best solution is to include both an alternative treatment and a traditional control group, but this of course adds cost and difficulty.

INTENT TO TREAT

An important principle of experimental research is that once students, classes, or schools are assigned to experimental and control conditions, they are considered part of those conditions, no matter what. For example, if a teacher is randomly assigned to the treatment group but then does not implement the treatment, his students are still assessed and included. Such subjects are called *intent-to-treat subjects* (see Begg, 2000). The reason it is important to include them is that dropping them introduces bias. Teachers who fail to implement may be less capable or less motivated than other teachers, so dropping them makes the remaining teachers look artificially good, on average. Similarly, if you randomly assign children to attend or not attend an after-school program, you will have to measure the children assigned to the after-school group even if they never show up, because to drop them will bias the experiment in favor of the after-school program (because the students who show up are probably more motivated than those who don't).

After the main intent-to-treat analysis, you can do an analysis of the subjects who did follow through, but this analysis will be given less importance.

Hawthorne effect: A tendency of subjects in an experimental group to exert outstanding efforts because they are conscious of being in an experiment, rather than because of the experimental treatments themselves.

The Savvy Researcher

Randomization

The Gremlin is a big fan of randomized experiments, but he knows that random assignment to experimental and control groups does not guarantee valid and meaningful research. Far from it.

Because randomized experiments are difficult to do, they often use small numbers of subjects. In educational research, small studies often suffer from teacher or class effects. For example, a researcher once told the Gremlin he was going to randomly assign 50 students to two classes. Class A, the experimental group, would be taught by Teacher A; Class B, the control group, would be taught by Teacher B. The Gremlin was not impressed. "If Teacher A is a very skilled teacher, the treatment will appear to be effective, even though it was the teacher, not the treatment, that made the difference," said the Gremlin. "Go find more teachers so you won't be confusing teachers and treatments."

The Gremlin is always on the lookout for randomized experiments that are very brief. In a year-long study of different ways of teaching sixth-grade mathematics, it's likely that the experimental and control groups both covered pretty much the same material. In a one-week study, however, focusing on a single skill (such as solving two-stage word problems), the Gremlin points out that it's likely that the experimental class will focus more on the skill being assessed in the experiment. It's also possible that the experimental and/or control group will do something unusual that they could keep up for a week but not for a month or a year, creating an artificial test of a treatment that may not have any meaning for educational practice or theory.

So use random assignment by all means, whenever possible. But as the Gremlin advises, don't assume that doing so will solve all problems!

PRETESTING

Many research methodologists advocate the use of experimental designs in which groups are randomly assigned to experimental and control groups, the experimental group receives the treatment, and then both groups are posttested (as in the Mexican American attitude film example discussed earlier). They argue that this design may be superior to pretest–posttest experimental designs, where randomly assigned groups are pretested, the treatment is applied to one group, and both groups are posttested. As noted earlier, pretests are not absolutely necessary when subjects are randomly assigned to treatments. But are they harmful or helpful?

Giving a pretest leaves open the possibility that the pretest may sensitize the subjects to the different treatments, leading to a false appearance that the treatment made a difference, when in fact the treatment would not have worked unless the pretest had also been given. This might occur in the study of the film designed to change attitudes toward Mexican Americans. Filling out a survey on their attitudes toward Mexican Americans just prior to seeing the film might make students especially sensitive to the film. However, the problem of sensitizing students to experimental treatments is rare in educational research, especially research on achievement. It is hard to see why a spelling pretest would sensitize students to one form of spelling instruction or another. In the case of the Mexican American attitude film, the students were not told that the purpose of the film was to change their attitudes, but a pretest might have tipped them off to it. It would be difficult to teach spelling, however, without students being aware of your primary objective.

While the dangers of pretesting in educational research are usually minimal, the dangers of not pretesting are great. What if the experimental and control groups are initially equal, but a few students leave school before the end of the project or refuse to fill out a valid posttest? Getting 100 percent of the data is rare in educational research. If any students are lost, we no longer know whether we have equivalent groups, and we have no way to do anything about it; the result is that the data derived are of limited usefulness.

Even more important, when pretests are given, they make it possible to use a common statistical method called **analysis of covariance** (see Chapter 14) to compare group means. In analysis of covariance (or equivalent procedures), scores on an outcome measure are adjusted for scores on some number of **covariates,** or control variables, such as pretests. Analysis of covariance can make groups that are somewhat different on a pretest effectively equivalent for statistical analysis.

For example, let's say that in an experiment with two treatments, we find that despite random assignment, students in Treatment A have a pretest mean of 6.4, while students in Treatment B have a pretest mean of 6.0. Of course, even if the treatments had no effect, we would expect students in Treatment A to score higher than those in Treatment B on the posttest because they started higher. Analysis of covariance would adjust the posttest scores to correct for this. Also, analysis of covariance usually increases **statistical power,** which is the ability of a statistic to find significant differences, if differences truly exist, and avoid a false negative error (see Chapter 14).

The advantage of analysis of covariance and related statistical procedures makes pretesting highly desirable in most educational research, particularly research on academic achievement and other variables likely to be correlated with achievement. So much of any test score is explained by student ability and past achievement that treatment effects are almost always small in relation to student-to-student differences. If these differences are not controlled for using analysis of covariance or a similar procedure, treatments that are in fact effective will often appear ineffective—a serious and common false negative error.

analysis of covariance: A statistical method that compares two or more group means after adjustment for some control variable or covariate (such as pretest) to see if any differences between the adjusted means are statistically significant.

covariate: A control variable used in analysis of covariance or multiple regression analysis to adjust other values.

statistical power: The ability of a statistical analysis to avoid getting a false negative error.

Here is the "bottom line": Unless you are worried about a pretest sensitizing students to the treatment, always give a pretest in a randomized experiment. You'll be glad you did (see Allison, 1995; Allison et al., 1997). For discussion of a related topic, see What If Pretests Are Not Equal in Different Treatment Groups? in Chapter 3.

Experiments with More Than Two Treatments

factorial design: An experimental comparison design in which treatments or other variables are analyzed as levels of one or more factors.

factor: A variable hypothesized to affect or cause another variable or variables; an independent variable.

continuous variable: A variable (such as age, test score, or height) that can take on a wide or infinite number of values.

dichotomous variable: A categorical variable (such as sex, on or off task, experimental control) that can take on only two values.

Of course, it is possible to compare more than two different treatments in an experimental comparison. For example, a researcher might randomly assign 90 students to three mathematics classes: one that takes weekly tests and gets daily homework, one that gets daily homework but no tests, and one that gets neither homework nor tests (see Figure 2.2). In this design, called a 3 × 1 experiment, we might pretest all students on their mathematics achievement, implement the treatments for several weeks, and then posttest. We would then compare the three average achievement levels on the posttests, statistically controlling for the pretests.

A **factorial design** is another type of experimental comparison design involving more than two treatments. In a factorial design, treatments may be organized in such a way that they share factors with other treatments. A **factor** is a variable that may take on a small number of values or categories. Examples of factors might include sex (male versus female), race/ethnicity (African American versus white versus Asian American), or type of school (private versus public). **Continuous variables**—such as achievement, age, and attitude (which can take on many values)—can be made into factors by establishing ranges of values for each level of the factor. For example, intelligence quotient (IQ) could be a factor with three levels: low (below 85), average (86 to 115), and high (above 115). Other variables could be dichotomized (reduced to two levels by splitting subjects into a high group [at or above the median] and a low group [below the median]). A **dichotomous variable** is a categorical variable that can take on only two values. For example, sex has only two values: male and female.

FIGURE 2.2

Hypothetical 3 × 1 Experimental Comparison

Homework + Tests	Homework Only	Control
N = 30	N = 30	N = 30

FIGURE 2.3

Hypothetical 2 × 2 Experimental Comparison

	Tests	No Tests
Homework	Homework + Tests	Homework Only
No Homework	Tests Only	Control

Experimental treatments may be seen as factors. For example, if the study involving homework and tests had four groups, it might have used a 2 × 2 factorial design, as depicted in Figure 2.3. The factors in the experiment depicted in this figure are homework (homework versus no homework) and tests (tests versus no tests). The factorial design has more statistical power (that is, smaller differences between means will be statistically significant) and produces more information than would a comparison of the same four treatments in a 4 × 1 analysis. A 2 × 2 analysis of variance or analysis of covariance for the study diagrammed in Figure 2.3 would produce a statistic for a homework factor, one for a test factor, and one for a homework-by-test interaction. Some of the possible outcomes of this factorial study are shown in Figure 2.4.

INTERACTIONS IN FACTORIAL EXPERIMENTS

An **interaction** describes a relationship between two factors in which a certain combination of the factors produces a result that is not simply the sum of the factors' **main effects.** For example, it has long been known that motivation is a product of the value of success and the probability of success. Imagine that you were told you could earn $1,000 if you won a tennis match. You would be highly motivated to practice and get in shape for the match, right? But what if you found out your opponent was tennis champion Venus Williams? Your probability of success would be zero (unless you're her sister, Serena), so your motivation to get in shape would be low. Motivation is high only when both value and probability are not zero. You would be more motivated by a $10 prize playing against someone you might beat than a large prize playing against the Williams sisters. This is therefore an example of an interaction: A combination of value and probability produces much more motivation than either factor by itself. This is illustrated in Table 2.1 on page 39.

interaction: An effect on a dependent (outcome) variable of a combination of two or more factors or independent variables that is not simply the sum of the separate effects of the variables

main effect: A simple effect of a factor or independent variable on a dependent (outcome) variable.

Figure 2.4 depicts some of the main effects and interactions that could have been seen in the 2 × 2 factorial experiment on homework and tests. A main effect indicates that on average, subjects who were at one level of a factor (e.g., homework versus no homework) scored differently than subjects at another level of the same

FIGURE 2.4

Some Possible Outcomes of a 2 × 2 Factorial Experiment

	Homework	No Homework
Tests	Hi	Lo
No Tests	Hi	Lo

Homework Main Effect
No Effect of Tests
No Interactions

	Homework	No Homework
Tests	Hi	Hi
No Tests	Lo	Lo

Tests Main Effect
No Effect of Homework
No Interactions

	Homework	No Homework
Tests	Hi	Lo
No Tests	Lo	Hi

No Main Effect
Homework by Tests
Interaction

	Homework	No Homework
Tests	Very Hi	Hi
No Tests	Hi	Hi

Two Main Effects
Homework by Tests
Interaction

factor, regardless of their scores on other factors (e.g., tests versus no tests). In Figure 2.4, a homework main effect would be observed if, on average, the students in the homework-plus-tests group and the homework-only group learned more than students in the tests-only group and the no-homework, no-tests group.

ORDINAL VERSUS DISORDINAL INTERACTIONS

Interactions are usually seen in factorial studies in which subjects in one cell score much better (or much worse) than would have been expected based on their respective factors. For example, in the homework and tests study, it is possible that either daily homework or regular tests would have a small positive effect on achievement, but a combination of the two would have a strong positive effect on achievement.

TABLE 2.1

Example of an Interaction: Probability of Success and Incentive in Tennis

This is an example of an interaction because the effect on motivation of the combination of probability of success and incentive is much greater than that of either factor by itself.

Situation	Probability of Success	Incentive	Motivation to Get in Shape
1. You vs. Venus Williams for no prize	0.00	0	Are you kidding?
2. You vs. Venus Williams for $1,000 prize	0.00	$1,000	Big deal. I'm never going to see that money!
3. You vs. a tennis player equal to you for no prize	0.50	0	Might be fun, but I'm not working that hard.
4. You vs. a tennis player equal to you for $1,000 prize	0.50	$1,000	Hand me my sneakers!

disordinal interaction: An interaction between treatment and other variables in which the rank order of treatment groups depends on other variables.

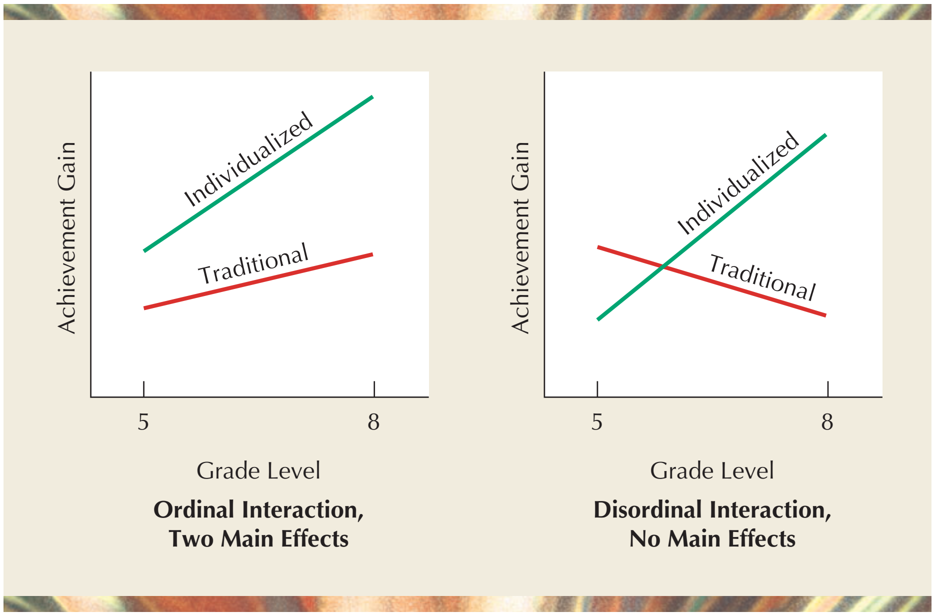
ordinal interaction: An interaction between treatment and other variables in which the rank order of the treatment groups does not depend on the other variables.

(This is shown in the lower-righthand table of Figure 2.4.) That is, homework and tests would have an interactive effect on achievement; they work better together than they work separately (analogous to the joint effect of value and probability in the tennis example).

Interactions can take many forms, and their forms have considerable bearing on their interpretations. Let's say we compared traditional instruction to an individualized instruction program in fifth- and eighth-grade classes and found that traditional instruction was better than individualized instruction in the fifth grade but the opposite was true in the eighth grade. This is called a **disordinal interaction** because the rank order of the treatments depends on values of the other variable. In this case, opposite results are obtained for type of instruction, depending on the grade level involved. Another possible outcome might have been that the individualized instruction program was a little better than traditional instruction for fifth-graders but much better than traditional instruction for eighth-graders. This is called an **ordinal interaction**; the results are larger at one grade level than at the

FIGURE 2.5

Ordinal and Disordinal Interactions



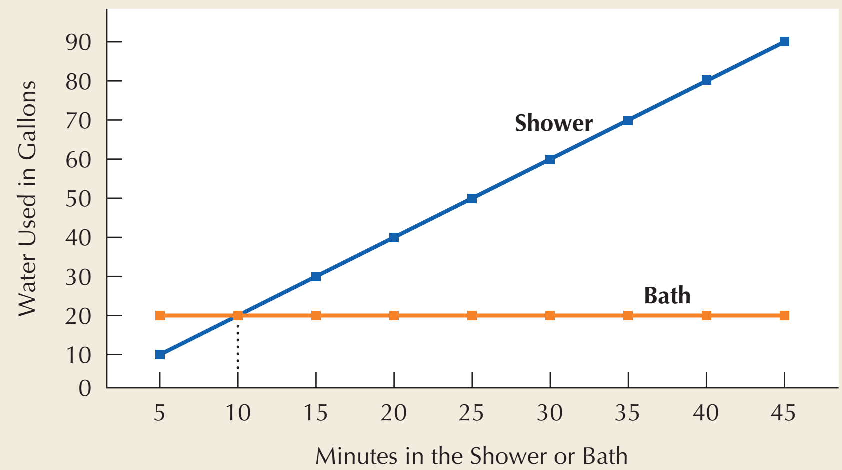
other, but the rank order of the treatments is the same at each grade level. (That is, the individualized treatment is higher at both grade levels.) Ordinal and disordinal interactions are diagrammed in Figure 2.5.

The meaning of an interaction depends to a considerable degree on whether the interaction is ordinal or disordinal. In a disordinal interaction, main effects for one factor depend completely on the other factor. In the disordinal interaction depicted in Figure 2.5, we cannot talk about the effects of individualized versus traditional instruction in general, because the effects depend totally on grade level. In contrast, in a case of an ordinal interaction, main effects are interpretable. In the example discussed earlier, individualized instruction is better than traditional instruction at both grade levels. However, the interaction does require caution about assuming that individualized instruction will be more effective than traditional instruction for, say, third-graders, because they fall outside the range we studied. We could make no assumptions about third-graders until we had replicated the study with third-graders.

As another example of interactions, consider an experiment to determine whether people use more water in showers or baths. If you have a combination bathtub and shower, you can do this experiment for yourself. Run a bath to your usual level, and mark that level with a grease pencil. Next time, take a shower with the drain plugged. Did you use more water or less? Clearly, the answer depends on how long you stay in the shower. As shown in Figure 2.6, the amount of water used in a bath is not affected by time, but the amount used in a shower depends totally on time. That is, there is an interaction between treatment (shower/bath) and time. A legitimate answer to our question about which uses more water is “It depends.”

FIGURE 2.6

The Shower/Bath Experiment: Time-by-Treatment Interactions



However, in our particular bathtub, water use is equal for showers and baths at 10 minutes. What if we found out that most people's showers are much longer than 10 minutes? Then we would know more than just "it depends." We could predict that for people on average, showers use more water.

The purpose of this example is to illustrate that we can learn much more from an interaction than "it depends." The form of the interaction often tells us conditions under which one treatment might be better than another as well as conditions under which the opposite might be the case.

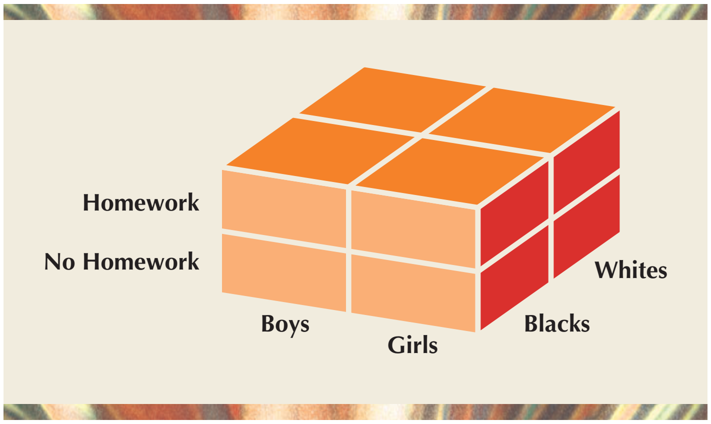
"After our research methods class, would you like to test the interaction effects of dinner and a movie?"

FACTORIAL DESIGNS WITH MORE THAN TWO FACTORS

A factorial design may have any number of factors, including a mix of treatments (to which subjects are randomly assigned by the experimenter) and factors over which the researcher has no experimental control. For example, a researcher might hypothesize different effects of homework for boys and girls and for African American and white students. She might set up a $2 \times 2 \times 2$ factorial design, as in Figure 2.7 on page 42. If such a design were used, the researcher would probably stratify on sex

FIGURE 2.7

Hypothetical $2 \times 2 \times 2$ Experimental Comparison



and race in making her random assignments to treatments (homework versus no homework) to make sure that boys, girls, African Americans, and whites are approximately equally distributed among the various cells. A three-factor experiment of this type would produce three main effects; three two-way interactions (homework by sex, homework by race, race by sex) and one three-way interaction (homework by sex by race).

Chapter 14 describes statistical procedures for factorial designs as well as for analysis of variance and analysis of covariance.

Alternatives to Random Assignment of Individuals

In evaluating the effect of experimental treatments, there is no design as powerful and conclusive as random assignment of individuals to experimental and control groups. This type of design virtually rules out selection bias as a source of error. Yet in educational research, random assignment of individual students is very difficult to achieve. Because the purpose of schooling is to educate and socialize students, not to provide a laboratory for researchers, schools are often less than enthusiastic about disrupting class assignments during the schoolyear and are no more positively inclined toward making permanent class assignments on a random basis.

When random assignment can be used, it is often for a short time or with a small group. For example, it may not be difficult to get a teacher to divide his or her class into two randomly assigned groups for a week or two, but for many kinds of research, these groups would be too small and the time period too brief for a meaningful study. Furthermore, a randomly assigned group is itself an innovation in schools, where ability grouping, student course selection, and other systematic assignments are the norm. A randomly chosen group of students who do not know one another may be so different from the typical classroom setting that research

Research with Class

Testing Effectiveness

Ed Decatur teaches fifth grade. He has always been a believer in field trips, especially to local museums, and he has read about the importance of experiential education (e.g., Joyce, Weil, & Calhoun, 2004). However, his principal, Ms. Chacon, is concerned about the expense and the children's time away from school. Ed proposes a study to test the effects of museum visits.

Ed randomly assigns his 24 students to two conditions: Half of the children will go to the natural history museum as part of a science unit on dinosaurs, and half will go to the museum of civilization as part of a history unit on ancient Egypt. Otherwise, he teaches each subject exactly as he always has.

At the end of the three-week units, he gives his usual end-of-unit tests. His hypothesis is that the children randomly assigned to the dinosaur trip will do better on the dinosaur test, while those who visited the Egyptian exhibit will do better on the Egypt test. The results partly support his hypothesis. There are no differences on the multiple-choice part of his tests, but the children who went to the museums did markedly better on the essay questions. Based on this research, Ms. Chacon encouraged Ed to continue his museum visits.

with such a group may have limited generalizability. In fact, many situations that allow for random assignment may already be so unusual that results from that situation may be difficult to apply to other settings.

RANDOM ASSIGNMENT OF CLASSES, SCHOOLS, AND TEACHERS

One practical procedure for experiments in schools is random assignment of classes (or schools) instead of students. Consider the study of the effects of homework and weekly tests on student achievement, described earlier. In the example, students in three classes were mixed up and randomly assigned to new classes, each implementing one of the three conditions. In that design, teacher effects would be completely confounded with treatment effects unless the teacher were rotated across classes. But the rotation itself would introduce practical as well as experimental design problems.

A great deal of research suggests children benefit from experiential activities. How would you know whether your students might benefit from such an activity? How could you determine how much they benefit from it?

As an alternative to this procedure, *classes* could be randomly assigned to the three treatments. For example, the researcher might solicit six fourth-, six fifth- and six sixth-grade teachers across two schools (a total of eighteen classes). She might assign classes to each condition within grades and within schools so that one intact class is assigned to each treatment at each grade level in each school. All classes are pretested, the treatments are implemented, and the students are posttested. Analysis of covariance is used.

This study has many advantages over the three-class version with individual random assignment. The larger number of teachers (six per experimental treatment group) makes it unlikely that teacher effects will be confounded with treatment effects. Rotation is unnecessary, and potential **school effects** are neutralized because there are three teachers in each school in each treatment. Random assignment of classes makes substantial pretest differences unlikely, but the analysis of covariance is capable of adjusting for any small differences that do exist. There are six times as many subjects in each group, making a false negative error less likely. The only real drawback to this design is that there are six times as many data to deal with and six times as many teachers and classes to monitor. When random assignment is done at the class level, at least five classes should be included in each treatment to reduce the chance of false treatment effects because of a peculiar class or teacher. As noted earlier, a quantitative methodologist would ask for 40 to 50 classes in such an experiment, so that analyses could be done at the class level, rather than the individual student level. This would eliminate the problem of potential teacher and class effects, but it is a larger study than most researchers could afford. Still, a

school effects: The effects on students or teachers of being in a particular school.

researcher with a modest budget can do a very good unbiased study with five classes per treatment and analysis at the student level.

DELAYED TREATMENT CONTROL GROUP DESIGNS

A serious recruitment problem in many experimental comparison designs in schools is that no one wants to be in the control group. Assuming that the experimental treatment is attractive, teachers may be reluctant to participate in a randomized experiment knowing that they have a 50-50 chance of getting nothing.

A solution to this problem is the **delayed treatment control group design.** In this design, teachers are told that they will all receive the experimental treatment, but some (determined at random) will get it right away and some at the end of the experiment. The delayed treatment group serves as a control group during the experiment. For example, imagine that you want to study the learning outcomes of a month-long middle school geography unit that uses simulation games. The simulation games are a lot of fun, and the teachers you approach are all eager to try them and don't want to be left out. You might get all the teachers who want to participate to agree to be randomly assigned to use the simulation games right way or to wait a month. This will add to the experiment the cost of the additional treatment materials and training, but it will also make the experiment more attractive to potential subjects, who will more likely do a good job of helping you collect data, do questionnaires, and so on. Use of delayed treatment control helps ensure that individuals who participate in the control group are, like those in the experimental group, willing to have the experimental treatment(s) applied to them.

WITHIN-TEACHER RANDOM ASSIGNMENT

In schools with departmentalization, where teachers have more than one class in the same subject, it is often possible to have teachers serve as their own controls by randomly assigning two or more of their classes to experimental and control (or Treatment 1 and Treatment 2) conditions. A very good study can be done in such circumstances with as few as three teachers, where each teaches at least two experimental and two control classes. A smaller number of teachers can be used because in this design, we need not be very concerned about teacher effects. We do still want at least three classes in each treatment condition, however, to reduce the chance that treatment effects are due to peculiarities of a single class.

When randomly assigning teachers' classes to different treatments (especially when classes are ability grouped), stratifying is critical. To do this, we would assign classes to treatments stratifying on teacher and on average class achievement in the same way that we assigned students to treatments stratifying on sex and achievement level. Consider a study in which the researcher wants to find out if weekly certificates for the most improved student will increase the motivation and achievement of all students. The researcher obtains the cooperation of three eighth-grade

delayed treatment control group design: A design with a control group that will receive the experimental treatment later, after the study is over.

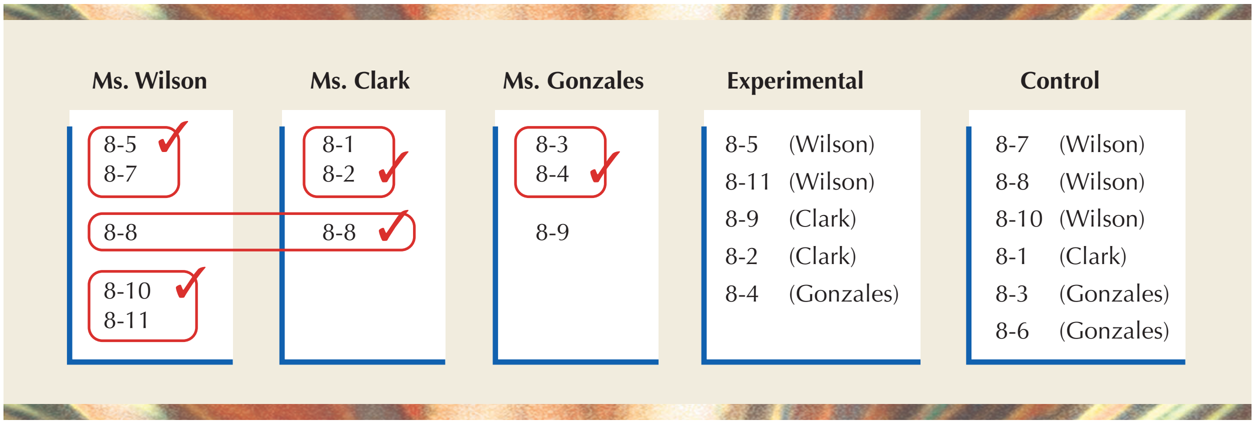


FIGURE 2.8

Example of Random Assignment of Classes

Note: A check (✔) indicates that the student was randomly assigned to the experimental group.

mathematics teachers in a middle school. One teacher, Ms. Wilson, volunteers all five of her classes. Mr. Clark has two classes for the talented and gifted that he does not want to be involved in the study, so he volunteers his other three classes. Ms. Gonzales teaches only three eighth-grade classes, which she volunteers for the study. Math classes are grouped by ability in this school. Based on the class averages on a standardized test, the classes are ranked from 8-1 (the highest-scoring eighth-grade class) to 8-11 (the lowest-scoring class). Each teacher's classes are listed in Figure 2.8. The researcher puts the classes into comparable pairs for random assignment, so that no matter how the coin flips go, the two groups will be about equal. Figure 2.8 illustrates the random assignment. The pairs are circled. Note that the middle-scoring class, 8-6, is left out at first. Because it is the middle class, it will make no difference to which group it is assigned.

To assign the classes, the researcher flips a coin to decide which class will be part of the experimental group and which will be part of the control group. That is, he flips a coin between 8-5 and 8-7, 8-8 and 8-9, 8-10 and 8-11, and so on. There need not be an even number of classes to assign. Note that Ms. Gonzales's 8-6 class, which was not initially paired because of the odd number of classes, is assigned to the control group on the basis of a coin flip. Again, this will not upset the comparability of the experimental and control groups, because 8-6 is an average-achieving class. The results of the coin flips and assignments to conditions are shown in the last two columns of Figure 2.8.

This random assignment makes the two groups comparable and ensures that each teacher has some experimental and some control classes, as evenly balanced as possible. By the vagaries of random assignment, the average rank of the experimental group (6.2) is somewhat higher than that of the control group (5.8). If the actual pretest scores reflect this difference, this will be well within the range where use of analysis of covariance can make the groups statistically equivalent. Once the groups are assigned, the study can proceed in the same way as any experimental comparison.

There are some limitations to the use of within-teacher random assignment of classes. This design cannot be used when there is a chance that the teachers will have trouble keeping the experimental and control treatments separate. In the two examples discussed in this section, this was not a problem; teachers could easily and reliably give homework and/or tests in some classes but not others or certificates in some classes but not others. However, consider a study in which the experimental treatment involves training teachers to ask questions that require students to think. It might be difficult for the same teacher to reliably ask mostly thinking questions in one class and mostly factual questions in another. If no differences were found between the treatments, the Gremlin might argue that the failure to find differences is due to the fact that the skills the teacher learned were also used in the control classes. In any study in which teachers serve as their own controls, it is particularly important to systematically observe the classes in order to verify that the treatments were reliably implemented (see Chapter 9).

Within-teacher random assignment can also create another more general problem. If teachers guess the researcher's hypothesis about which treatment is best, they may help that treatment be best in ways other than simply implementing the prescribed procedures, such as by giving favored treatment to the experimental class. The researcher can reduce this unwanted "help" by being very clear with the teachers that the purpose of the experiment is not to prove a hypothesis but to give it a fair test, by refraining from making his or her hopes too obvious, and by carefully monitoring the classes both during project implementation and during testing. This problem is greater when there is a control group that is obviously a control group than when there are two treatments being compared. In some research, researchers write teachers' manuals for the control group that essentially formalize what the teachers were already doing. By presenting this method as an alternative treatment, rather than an untreated control group, and by presenting the study as a comparison of two interesting methods, the researcher can be more confident that each treatment will get the teachers' best efforts.

Another limitation of within-teacher random assignment is that it is hard to use experimental designs that involve more than two treatments. It is usually too much to expect a teacher to implement three or four different treatments in different classes.

Although the problems and limitations of within-teacher random assignment should be carefully considered, this design does largely solve the very serious problem of confounding of teacher effects and treatment effects. This makes it a very useful design in departmentalized schools. Table 2.2 on page 48 summarizes the advantages and disadvantages of the four approaches to random assignment discussed in this chapter.

TABLE 2.2

Approaches to Random Assignment in Education

Approach	Advantages	Disadvantages
1. Random assignment of individuals	Requires fewest subjects.	Educators resist assigning students to classes or schools at random. May create artificial groupings.
2. Stratified random assignment of individuals or groups	Increases chances of equality on key factors. Ensures equal numbers in subgroups.	Requires somewhat larger sample size.
3. Random assignment of classes, schools, or teachers	Much more acceptable to educators than individual assignment.	Requires many more students than individual assignment.
4. Within-teacher random assignments (same teacher teaches experimental and control classes)	Very practical, especially in middle and high schools. Controls for teacher effects.	Teachers may have trouble keeping treatments separate.

EXAMPLE OF AN EXPERIMENT

Appendix 7 on pages 314–335 shows an example of an article reporting an experimental study by Baker, Gersten, and Keating (2000). It evaluates a volunteer tutoring program called SMART for low-achieving children in grades K–2. Look at the Methods section. It describes how children were first matched on a test of rapid letter naming and then assigned at random to be tutored or to serve in a control group. They were then tutored for six months, two days a week. Table 1 of the study shows that the SMART and comparison students were very well matched at pretest. At posttest, the children who experienced SMART scored significantly better on most measures (see Table 3).

Because of the use of random assignment to conditions, the researchers could be certain that the only important difference between the two groups was the tutoring itself. This type of study provides convincing evidence for the effectiveness of a treatment, because it eliminates virtually all other explanations for the findings other than that the treatment made the difference.

RESEARCH NAVIGATOR

Key Terms

Activity

If you have access to Research Navigator, located on the MyLabSchool website (www.mylabschool.com), enter the following keywords to find articles related to the content of this chapter:

- Randomized experiments
- Analysis of variance
- Analysis of covariance
- Control groups
- Interaction effects

EXERCISES

1. A researcher proposed to test two three-week social studies modules with three teachers, each of whom had three classes. Two teachers volunteered to participate. Each teacher was randomly assigned to implement one of the modules. The classes taught by the third teacher, who had declined to participate, served as a control group. What are the problems with this method of assignment?

2. Describe three alternative means of assigning students and teachers to treatments in a study such as the one described in exercise 1. Suppose that a researcher wants to assess the impact of a physical education game on arm strength with a group of 70 boys and girls aged 5 to 10. Describe a method of stratified random assignment for this study.

3. A researcher found the following results in a randomized factorial experiment regarding the impact of teacher recognition and higher salary on teacher morale:

   Group 1 Teacher recognition, high salary: teacher morale high

   Group 2 High salary, no teacher recognition: teacher morale low

   Group 3 Teacher recognition, low salary: teacher morale low

   Group 4 No teacher recognition, low salary: teacher morale low

   How would you describe this result in terms of main effects and interactions? How would you describe the meaning of this result to the teachers involved?

4. Be the Gremlin. Evaluate the study described in Research with Class (see page 42). Is Ed Decatur's experiment free from bias? How confident should he be in his conclusions? Why?

FURTHER READING

Learn more about the concepts discussed in this chapter by reviewing some of the research cited.

Randomized Experiments

Bloom, H. S. (Ed.). (2005). *Learning more from social experiments: Evolving analytic approaches.* New York: Russell Sage Foundation.

Christensen, L. (2001). *Experimental methodology.* Boston: Allyn & Bacon.

Martin, D. W. (2004). *Doing psychology experiments* (6th ed.). Belmont, CA: Wadsworth.

Mosteller, F., & Boruch, R. (Eds.). (2002). *Evidence matters: Randomized trials in education research.* Washington, DC: Brookings Institution Press.

Phye, G. D., Robinson, D. H., & Levin, J. (2005). *Empirical methods for evaluating educational interventions.* Oxford, England: Elsevier.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* New York: Houghton-Mifflin.

Towne, L., & Hilton, M. (Eds.). (2004). *Implementing randomized field trials in education: Report of a workshop.* Washington, DC: National Academies Press.