# 13 Experimental Research




**OBJECTIVES** Studying this chapter should enable you to:

- Describe briefly the purpose of experimental research.
- Describe the basic steps involved in conducting an experiment.
- Describe two ways in which experimental research differs from other forms of educational research.
- Explain the difference between random assignment and random selection and the importance of each.
- Explain what is meant by the phrase "manipulation of variables" and describe three ways in which such manipulation can occur.
- Distinguish between examples of weak and strong experimental designs and draw diagrams of such designs.
- Identify various threats to internal validity associated with different experimental designs.
- Explain three ways in which various threats to internal validity in experimental research can be controlled.
- Explain how matching can be used to equate groups in experimental studies.
- Describe briefly the purpose of factorial and counterbalanced designs and draw diagrams of such designs.
- Describe briefly the purpose of a time-series design and draw a diagram of this design.
- Describe briefly how to assess probable threats to internal validity in an experimental study.
- Recognize an experimental study when you see one in the literature.

**INTERACTIVE AND APPLIED LEARNING** After, or while, reading this chapter:

Go to McGraw Hill Connect® to:

- Learn More About What Constitutes an Experiment

Go to your online Student Mastery Activities book to do the following activities:

- Activity 13.1: Group Experimental Research Questions
- Activity 13.2: Designing an Experiment
- Activity 13.3: Characteristics of Experimental Research
- Activity 13.4: Random Selections vs. Random Assignment

Does team-teaching improve the achievement of students in high school social studies classes? Alicia Johnson, the principal of a large high school in Minneapolis, Minnesota, having heard encouraging remarks about the idea at a recent educational conference, wants to find out. Accordingly, she asks some of her eleventh-grade world history teachers to participate in an experiment. Three teachers are to combine their classes into one large group. These teachers are to work as a team, sharing the planning, teaching, and evaluation of these students. Three other teachers are assigned to teach a class in the same subject individually, with the usual arrangement of one teacher per class. The students selected to participate are similar in ability, and the teachers will teach at the same time, using the same curriculum. All are to use the same standardized tests and other assessment instruments, including written tests prepared jointly by the six teachers. Periodically during the semester, Ms. Johnson will compare the scores of the two groups of students on these tests.

This is an example of an experiment—the comparison of a treatment group with a nontreatment group. In this chapter, you will learn about various procedures that researchers use to carry out such experiments, as well as how they try to ensure that it is the experimental treatment rather than some uncontrolled variable that causes the changes in achievement.

Experimental research is one of the most powerful research methodologies that researchers can use. Of the many types of research that might be used, the experiment is the best way to establish cause-and-effect relationships among variables. Yet experiments are not always easy to conduct. In this chapter, we will show you both the power of, and the problems involved in, conducting experiments.

## The Uniqueness of Experimental Research

Of all the research methodologies described in this book, **experimental research** is unique in two very important respects: It is the only type of research that directly attempts to influence a particular variable, and when properly applied, it is the best type for testing hypotheses about cause-and-effect relationships. In an experimental study, researchers look at the effect(s) of at least one independent variable on one or more dependent variables. The **independent variable** in experimental research is also frequently referred to as the **experimental**, or **treatment**, **variable**. The **dependent variable**, also known as the **criterion**, or **outcome**, **variable**, refers to the results or outcomes of the study.

The major characteristic of experimental research that distinguishes it from all other types of research is that researchers *manipulate* the independent variable. They decide the nature of the treatment (i.e., what is going to happen to the subjects of the study), to whom it is to be applied, and to what extent. Independent variables frequently manipulated in educational research include methods of instruction, types of assignment, learning materials, rewards given to students, and types of questions asked by teachers. Dependent variables that are frequently studied include achievement, interest in a subject, attention span, motivation, and attitudes toward school.

After the treatment has been administered for an appropriate length of time, researchers observe or measure the groups receiving different treatments (by means of a posttest of some sort) to see if they differ. Another way of saying this is that researchers want to see whether the treatment made a difference. If the average scores of the groups on the posttest do differ and researchers cannot find any sensible alternative explanations for this difference, they can conclude that the treatment did have an effect and is likely the cause of the difference.

Experimental research, therefore, enables researchers to go beyond description and prediction, beyond the identification of relationships, to at least a partial determination of what causes them. Correlational studies may demonstrate a strong relationship between socioeconomic level and academic achievement, for instance, but they cannot demonstrate that improving socioeconomic level will necessarily improve achievement. Only experimental research has this capability. Some actual examples of the kinds of experimental studies that have been conducted by educational researchers are:

- The effect of small classes on instruction.[1]
- The effect of early reading instruction on growth rates of kindergarteners with learning differences.[2]
- The use of intensive mentoring to help beginning teachers develop balanced instruction.[3]
- The effect of lotteries on Web survey response rates.[4]
- Introduction of a course on bullying into preservice teacher-training curriculum.[5]
- Using social stories to enhance the interpersonal conflict resolution skills of children with learning disabilities.[6]
- Improving the self-concept of students through the use of guided meditation and hypnosis.[7]

# Essential Characteristics of Experimental Research

The word **experiment** has a long and illustrious history in the annals of research. It has often been hailed as the most powerful method that exists for studying cause and effect. Its origins go back to the very beginnings of history when, for example, primeval humans first experimented with ways to produce fire. One can imagine countless trial-and-error attempts on their part before achieving success by sparking rocks or by spinning wooden spindles in dry leaves. Much of the success of modern science is due to carefully designed and meticulously implemented experiments.

The basic idea underlying all experimental research is really quite simple: Try something and systematically observe what happens. Formal experiments consist of two basic conditions. First, at least two (but often more) conditions or methods are *compared* to assess the effect(s) of particular conditions or "treatments" (the independent variable). Second, the independent variable is directly *manipulated* by the researcher. Change is planned for and deliberately manipulated to study its effect(s) on one or more outcomes (the dependent variable). Let us discuss some important characteristics of experimental research in a bit more detail.

## COMPARISON OF GROUPS

An experiment usually involves two groups of subjects, an experimental group and a control or a comparison group, although it is possible to conduct an experiment with only one group (by providing all treatments to the same subjects) or with three or more groups. The **experimental group** receives a treatment of some sort (such as a new textbook or a different method of teaching), while the **control group** receives no treatment (or the **comparison group** receives a different treatment). The control or the comparison group is crucially important in all experimental research, for it enables the researcher to determine whether the treatment has had an effect or whether one treatment is more effective than another.

Historically, a pure control group is one that receives no treatment at all. While this is often the case in medical or psychological research, it is rarely true in educational research. The control group almost always receives a different treatment of some sort. Some educational researchers, therefore, refer to comparison groups rather than to control groups.

Consider an example. Suppose a researcher wished to study the effectiveness of a new method of teaching science. He or she would have the students in the experimental group taught by the new method, but the students in the comparison group would continue to be taught by their teacher's usual method. The researcher would not administer the new method to the experimental group and have a control group *do nothing.* Any method of instruction would likely be more effective than no method at all!

## MANIPULATION OF THE INDEPENDENT VARIABLE

The second essential characteristic of all experiments is that the researcher actively *manipulates* the independent variables. What does this mean? Simply put, it means that the researcher deliberately and directly determines what forms the independent variable will take and then which group will get which form. For example, if the independent variable in a study is the amount of enthusiasm an instructor displays, a researcher might train two teachers to display different amounts of enthusiasm as they teach their classes.

Although many independent variables in education can be manipulated, many others cannot. Examples of independent variables that can be manipulated include teaching method, type of counseling, learning activities, assignments given, and materials used; examples of independent variables that cannot be manipulated include gender, ethnicity, age, and religious preference. Researchers can manipulate the kinds of learning activities to which students are exposed in a classroom, but they cannot manipulate, say, religious preference—that is, students cannot be "made into" Protestants, Catholics, Jews, or Muslims, for example, to serve the purposes of a study. To manipulate a variable, researchers must decide who is to get something and when, where, and how they will get it.

The independent variable in an experimental study may be established in several ways—either (a) one form of the variable versus another; (b) presence versus absence of a particular form; or (c) varying degrees of the same form. An example of (a) would be a study comparing the inquiry method with the lecture method of instruction in teaching chemistry. An example of (b) would be a study comparing the use of PowerPoint slides versus no PowerPoint slides in teaching statistics. An example of (c) would be a study comparing the effects of different specified amounts of teacher enthusiasm on student attitudes toward mathematics. In both (a) and (b), the variable (method) is clearly categorical. In (c), a variable that in actuality is quantitative (*degree* of enthusiasm) is treated as categorical (the effects of only specified *amounts* of enthusiasm will be studied) in order for the researcher to manipulate (i.e., to control for) the amount of enthusiasm.

## RANDOMIZATION

An important aspect of many experiments is the random assignment of subjects to groups. Although there are certain kinds of experiments in which random assignment is not possible, researchers try to use randomization whenever feasible. It is a crucial ingredient in the best kinds of experiments. Random assignment is similar, but not identical, to the concept of random selection we discussed in Chapter 6. **Random assignment** means that every individual who is participating in an experiment has an equal chance of being assigned to any of the experimental or control conditions being compared. **Random selection**, on the other hand, means that every member of a population has an equal chance of being selected to be a member of the sample. Under random assignment, each member of the sample is given a number (arbitrarily), and a table of random numbers (see Chapter 6) is then used to select the members of the experimental and control groups.

Three things should be noted about the random assignment of subjects to groups. First, it takes place before the experiment begins. Second, it is a *process* of assigning or distributing individuals to groups, not a result of such distribution. This means that you cannot look at two groups that have already been formed and be able to tell, just by looking, whether or not they were formed randomly. Third, the use of random assignment allows the researcher to form groups that, right at the beginning of the study, are *equivalent*—that is, they differ only by chance in any variables of interest. In other words, random assignment is intended to eliminate the threat of **extraneous**, or additional, **variables**—not only those of which researchers are aware but also those of which they are not aware—that might affect the outcome of the study. This is the beauty and the power of random assignment. It is one of the reasons why experiments are, in general, more effective than other types of research for assessing cause-and-effect relationships.

This last statement is tempered, of course, by the realization that groups formed through random assignment may still differ somewhat. Random assignment ensures only that groups are equivalent (or at least as equivalent as human beings can make them) at the beginning of an experiment.

Furthermore, random assignment is no guarantee of equivalent groups unless both groups are sufficiently large. No one would expect random assignment to result in equivalence if only five subjects were assigned to each group, for example. There are no rules for determining how large groups must be, but most researchers are uncomfortable relying on random assignment with fewer than 40 participants in each group.

## Control of Extraneous Variables

Researchers in an experimental study have an opportunity to exercise far more control than in most other forms of research. They determine the treatment (or treatments), select the sample, assign individuals to groups, decide which group will get the treatment, try to control other factors besides the treatment that might influence the outcome of the study, and then (finally) observe or measure the effect of the treatment on the groups when the treatment is completed.

In Chapter 9, we introduced the idea of internal validity and discussed several kinds of threats to internal validity. It is very important for researchers conducting an experimental study to do their best to **control** for—that is, to eliminate or to minimize the possible effect of—these threats. If researchers are unsure whether another variable might be the cause of a result observed in a study, they cannot be sure what the cause really is. For example, if a researcher attempted to compare the effects of two different methods of instruction on student attitudes toward history but did not make sure that the groups involved were equivalent in ability, then ability might be a possible alternative explanation (rather than the difference in methods) for any differences in attitudes of the groups found on a posttest.

In particular, researchers who conduct experimental studies try their best to control any and all subject characteristics that might affect the outcome of the study. They do this by ensuring that the two groups are as equivalent as possible on all variables other than the one or ones being studied (i.e., the independent variables).

How do researchers minimize or eliminate threats due to subject characteristics? Many ways exist. Here are some of the most common:

**Randomization:** As we mentioned before, if subjects can be randomly assigned to the various groups involved in an experimental study, researchers can assume that the groups are equivalent. This is the best way to ensure that the effects of one or more possible extraneous variables have been controlled.

**Holding certain variables constant:** The idea here is to eliminate the possible effects of a variable by removing it from the study. For example, if a researcher suspects that gender might influence the outcomes of a study, she could control for it by restricting the subjects of the study to females and by excluding all males. The variable of gender, in other words, is held constant. However, there is a cost involved (as there almost always is) for this control, as the generalizability of the results of the study are correspondingly reduced.

**Building the variable into the design:** This solution involves building the variable(s) into the study to assess their effects. It is the exact opposite of the previous idea. Using the preceding example, the researcher would include *both* females and males (as distinct groups) in the design of the study and then analyze the effects of *both* gender and method on outcomes.

**Matching:** Often pairs of subjects can be matched on certain variables of interest. If a researcher felt that age, for example, might affect the outcome of a study, he might endeavor to match students according to their ages and then assign one member of each pair (randomly if possible) to each of the comparison groups.

**Using subjects as their own controls:** When subjects are used as their own controls, their performance under both (or all) treatments is compared. Thus, the same students might be taught algebra units first by an inquiry method and later by a lecture method. Another example is the assessment of an individual's behavior during a period of time before and after a treatment is implemented to see whether changes in behavior occur.

**Using analysis of covariance:** As mentioned in Chapter 11, analysis of covariance can be used to equate groups statistically on the basis of a pretest or other variables. The posttest scores of the subjects in each group are then adjusted accordingly.

We will shortly show you a number of research designs that illustrate how several of these controls can be implemented in an experimental study.

## Group Designs in Experimental Research

The **design** of an experiment can take a variety of forms. Some of the designs we present in this section are better than others, however. Why "better"? Because of the various threats to internal validity identified in Chapter 9: Good designs control many of these threats, while poor designs control only a few. The quality of an experiment depends on how well the various threats to internal validity are controlled.

## POOR EXPERIMENTAL DESIGNS

Designs that are "weak" do not have built-in controls for threats to internal validity. In addition to the independent variable, there are a number of other plausible explanations for any outcomes that occur. As a result, any researcher who uses one of these designs has difficulty assessing the effectiveness of the independent variable.

**The One-Shot Case Study.** In the **one-shot case study design**, a single group is exposed to a treatment or event and a dependent variable is subsequently observed (measured) to assess the effect of the treatment. A diagram of this design is as follows:

**The One-Shot Case Study Design**

| *X* | *O* |
|---|---|
| Treatment | Observation (Dependent variable) |

The symbol *X* represents exposure of the group to the treatment of interest, while *O* refers to observation (measurement) of the dependent variable. The placement of the symbols from left to right indicates the order in time of *X* and *O*. As you can see, the treatment, *X,* comes before observation of the dependent variable, *O*.

Suppose a researcher wishes to see if a new textbook increases student interest in history. He uses the textbook (*X*) for a semester and then measures student interest (*O*) with an attitude scale. A diagram of this example is shown in Figure 13.1.

The most obvious weakness of this design is its absence of any control. The researcher has no way of knowing if the results obtained at *O* (as measured by the attitude scale) are due to treatment *X* (the textbook). The design does not provide for any comparison, so the researcher cannot compare the treatment results (as measured by the attitude scale) with the same group before using the new textbook, or with those of another group using a different textbook. Because the group has not been pretested in any way, the researcher knows nothing about what the group was like before using the text. Thus, he does not know whether the treatment had *any* effect at all. It is quite possible that the students who use the new textbook *will* indicate very favorable attitudes toward history. But the question remains, were these attitudes produced by the new textbook? Unfortunately, the one-shot case study does not help us answer this question. To remedy this design, a comparison could be made with another group of students who had the same course content presented in the regular textbook. (We shall show you just such a design shortly.) Fortunately, the flaws in the one-shot design are so well known that it is seldom used in educational research.

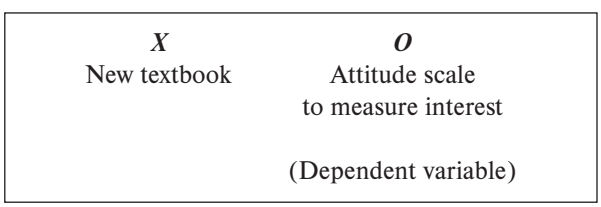



**Figure 13.1** *Example of a One-Shot Case Study Design*

**The One-Group Pretest-Posttest Design.** In the **one-group pretest-posttest design**, a single group is measured or observed not only after being exposed to a treatment of some sort, but also before. A diagram of this design is as follows:

**The One-Group Pretest-Posttest Design**

| *O* | *X* | *O* |
|---|---|---|
| Pretest | Treatment | Posttest |

Consider an example of this design. A principal wants to assess the effects of weekly counseling sessions on the attitudes of certain "hard-to-reach" students in her school. She asks the counselors in the program to meet once a week with these students for a period of 10 weeks, during which sessions the students are encouraged to express their feelings and concerns. She uses a 20-item scale to measure student attitudes toward school both immediately before and after the 10-week period. Figure 13.2 presents a diagram of the design of the study.

This design is better than the one-shot case study (the researcher at least knows whether any change occurred), but it is still weak. Nine uncontrolled-for threats to

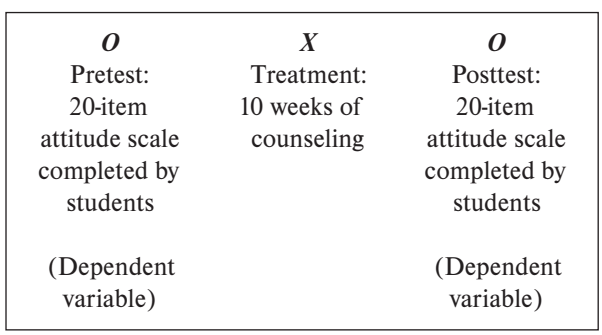



**Figure 13.2** *Example of a One-Group Pretest-Posttest Design*

internal validity exist that might also explain the results on the posttest. They are history, maturation, instrument decay, data collector characteristics, data collector bias, testing, statistical regression, attitude of subjects, and implementation. Any or all of these may influence the outcome of the study. The researcher would not know if any differences between the pretest and the posttest are due to the treatment or to one or more of these threats. To remedy this, a comparison group, which does not receive the treatment, could be added. Then if a change in attitude occurs between the pretest and the posttest, the researcher has reason to believe that it was caused by the treatment (symbolized by *X*).

**The Static-Group Comparison Design.** In the **static-group comparison design**, two already existing, or intact, groups are used. These are sometimes referred to as *static groups,* hence the name for the design. This design is sometimes called a **nonequivalent control group design**. A diagram of this design is as follows:

**The Static-Group Comparison Design**

| | |
|---|---|
| *X* | *O* |
| | *O* |

The dashed line indicates that the two groups being compared are already formed—that is, the subjects are not randomly assigned to the two groups. *X* symbolizes the experimental treatment. The blank space in the design indicates that the "control" group does not receive the experimental treatment; it may receive a different treatment or no treatment at all. The two *O*s are placed exactly vertical to each other, indicating that the observation or measurement of the two groups occurs at the same time.

Consider again the example used to illustrate the one-shot case study design. We could apply the static-group comparison design to this example. The researcher would (a) find two intact groups (two classes), (b) assign the new textbook (*X*) to one of the classes but have the other class use the regular textbook, and then (c) measure the degree of interest of all students in both classes at the same time (e.g., at the end of the semester). Figure 13.3 presents a diagram of this example.

Although this design provides better control over history, maturation, testing, and regression threats,* it is more vulnerable not only to mortality and location,† but also, more importantly, to the possibility of differential subject characteristics.

| | |
|---|---|
| ***X*** | ***O*** |
| New textbook | Attitude scale to measure interest |
| | ***O*** |
| Regular textbook | Attitude scale to measure interest |

**Figure 13.3** *Example of a Static-Group Comparison Design*

**The Static-Group Pretest-Posttest Design.** The **static-group pretest-posttest design** differs from the static-group comparison design only in that a pretest is given to both groups. A diagram for this design is as follows:

**The Static-Group Pretest-Posttest Design**

| | | |
|---|---|---|
| *O* | *X* | *O* |
| *O* | | *O* |

In analyzing the data, each individual's pretest score is subtracted from his or her posttest score, thus permitting analysis of "gain" or "change." While this provides better control of the subject characteristics threat (since it is the *change* in each student that is analyzed), the amount of gain often depends on initial performance; that is, the group scoring higher on the pretest is likely to improve more (or in some cases less), and thus subject characteristics still remains somewhat of a threat. Further, administering a pretest raises the possibility of a testing threat. In the event that the pretest is used to match groups, this design becomes the matching-only pretest-posttest control group design (see p. 269), a much more effective design.

## TRUE EXPERIMENTAL DESIGNS

The essential ingredient of a true experimental design is that subjects are randomly assigned to treatment groups. As discussed earlier, random assignment is a powerful technique for controlling the subject characteristics threat to internal validity, a major consideration in educational research.

*History and maturation remain possible threats because the researcher cannot be sure that the two groups have been exposed to the same extraneous events or have the same maturational processes.

†This is because the groups may differ in the number of subjects lost and/or in the kinds of resources provided.

**The Randomized Posttest-Only Control Group Design.** The **randomized posttest-only control group design** involves two groups, both of which are formed by random assignment. One group receives the experimental treatment while the other does not, and then both groups are posttested on the dependent variable. A diagram of this design is as follows:

**The Randomized Posttest-Only Control Group Design**

| | | | |
|---|---|---|---|
| Treatment group | *R* | *X* | *O* |
| Control group | *R* | *C* | *O* |

As before, the symbol *X* represents exposure to the treatment and *O* refers to the measurement of the dependent variable. *R* represents the random assignment of individuals to groups. *C* now represents the control group.

In this design, the control of certain threats is excellent. Through the use of random assignment, the threats of subject characteristics, maturation, and statistical regression are well controlled for. Because none of the subjects in the study are measured twice, testing is not a possible threat. This is perhaps the best of all designs to use in an experimental study, provided there are at least 40 subjects in each group.

There are, unfortunately, some threats to internal validity that are not controlled for by this design. The first is mortality. Because the two groups are similar, we might expect an equal dropout rate from each group. However, exposure to the treatment may cause more individuals in the experimental group to drop out (or stay in) than in the control group. This may result in the two groups becoming dissimilar in terms of their characteristics, which in turn may affect the results on the posttest. For this reason, researchers should always report how many subjects drop out of each group during an experiment. An attitudinal threat is possible. In addition, implementation, data collector bias, location, and history threats may exist. These threats can sometimes be controlled by appropriate modifications to this design.

As an example of this design, consider a hypothetical study in which a researcher investigates the effects of a series of sensitivity training workshops on faculty morale in a large high school district. The researcher randomly selects a sample of 100 teachers from all the teachers in the district. The researcher then (a) randomly assigns the teachers in the district to two groups; (b) exposes one group, but not the other, to the training; and then (c) measures the morale of each group using a questionnaire. Figure 13.4 presents a diagram of this hypothetical experiment.

Again we stress that it is important to keep clear the distinction between random selection and random assignment. Both involve the process of randomization, but for a different purpose. Random selection, you will recall, is intended to provide a representative sample. But it may or may not be accompanied by the random assignment of subjects to groups. Random assignment is intended to equate groups, and often is not accompanied by random selection.

**The Randomized Pretest-Posttest Control Group Design.** The **randomized pretest-posttest control group design** differs from the randomized posttest-only control group design solely in the use of a pretest. Two groups of subjects are used, with both groups being measured or observed twice. The first measurement serves as the pretest, the second as the posttest.

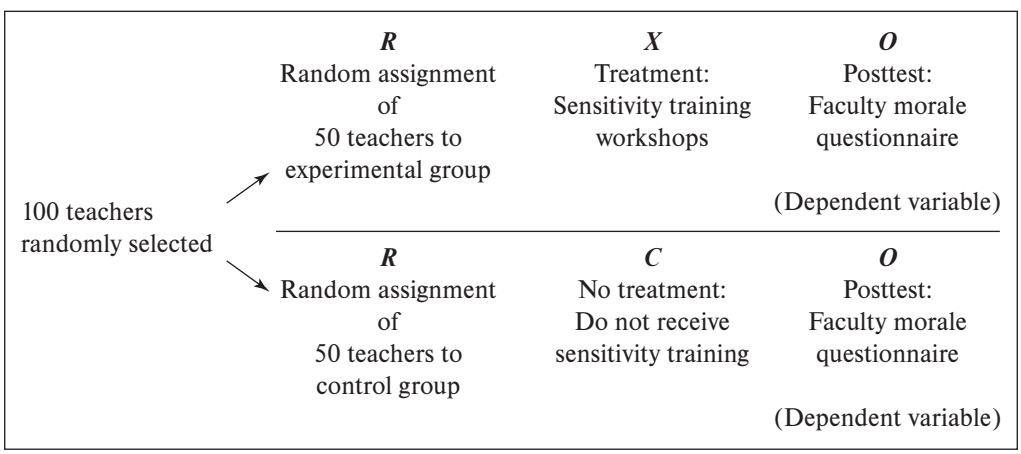



**Figure 13.4** *Example of a Randomized Posttest-Only Control Group Design*

Random assignment is used to form the groups. The measurements or observations are collected at the same time for both groups. A diagram of this design follows:

**The Randomized Pretest-Posttest Control Group Design**

| | | | | |
|---|---|---|---|---|
| Treatment group | *R* | *O* | *X* | *O* |
| Control group | *R* | *O* | *C* | *O* |

The use of the pretest raises the possibility of a **pretest treatment interaction** threat, since it may "alert" the members of the experimental group, thereby causing them to do better (or more poorly) on the posttest than the members of the control group. A trade-off is that it provides the researcher with a means of checking whether the two groups are really similar—that is, whether random assignment actually succeeded in making the groups equivalent. This is particularly desirable if the number in each group is small (less than 30). If the pretest shows that the groups are not equivalent, the researcher can seek to make them so by using one of the **matching designs** we will discuss shortly. A pretest is also necessary if the amount of change over time is to be assessed.

Let us illustrate this design by using our previous example involving the use of sensitivity workshops. Figure 13.5 presents a diagram of how this design would be used.

**The Randomized Solomon Four-Group Design.** The **randomized Solomon four-group design** is an attempt to eliminate the possible effect of a pretest. It involves random assignment of subjects to four groups, with two of the groups being pretested and two not. One of the pretested groups and one of the unpretested groups is exposed to the experimental treatment. All four groups are then posttested. A diagram of this design is as follows:

**The Randomized Solomon Four-Group Design**

| | | | | |
|---|---|---|---|---|
| Treatment group | *R* | *O* | *X* | *O* |
| Control group | *R* | *O* | *C* | *O* |
| Treatment group | *R* | | *X* | *O* |
| Control group | *R* | | *C* | *O* |

The randomized Solomon four-group design combines the pretest-posttest control group and posttest-only control group designs. The first two groups represent the pretest-posttest control group design, while the last two groups represent the posttest-only control group design. Figure 13.6 presents an example of the randomized Solomon four-group design.

The randomized Solomon four-group design provides the best control of the threats to internal validity that we have discussed. A weakness, however, is that it requires a large sample because subjects must be assigned to four groups. Furthermore, conducting a study involving four groups at the same time requires a considerable amount of energy and effort on the part of the researcher.

**Random Assignment with Matching.** In an attempt to increase the likelihood that the groups of subjects in an experiment will be equivalent, pairs of

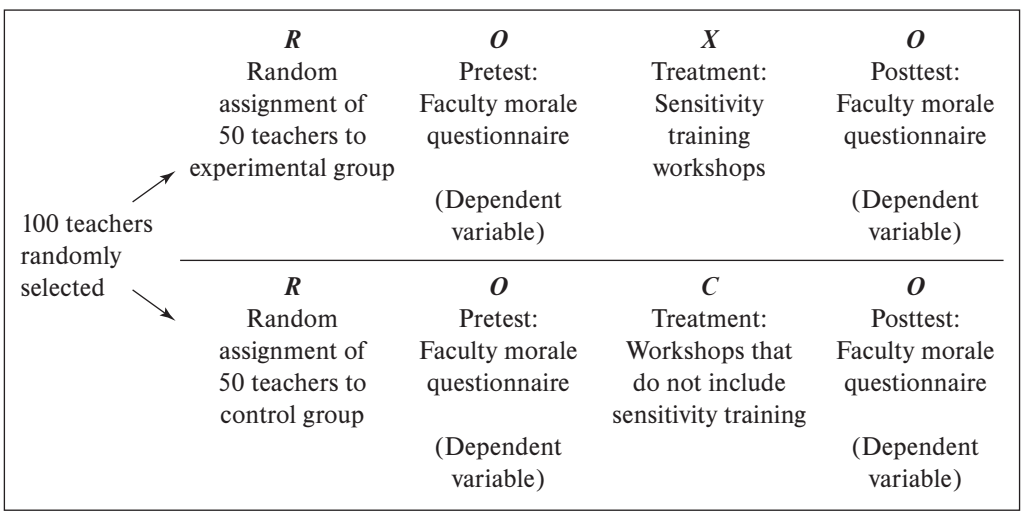



**Figure 13.5** *Example of a Randomized Pretest-Posttest Control Group Design*

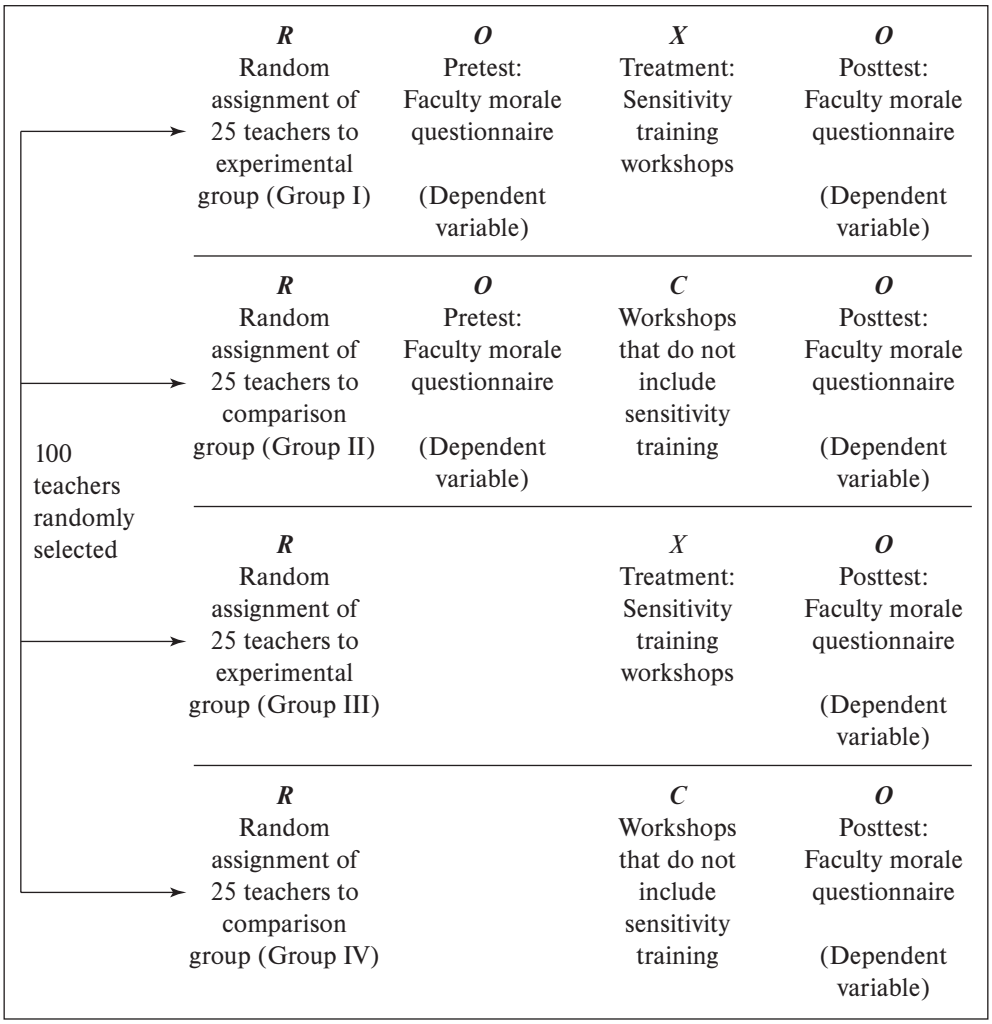



**Figure 13.6** *Example of a Randomized Solomon Four-Group Design*

individuals may be matched on certain variables. The choice of variables on which to match is based on previous research, theory, and/or the experience of the researcher. The members of each matched pair are then assigned to the experimental and control groups at random. This adaptation can be made to both the posttest-only control group design and the pretest-posttest control group design, although the latter is more common. Diagrams of these designs are:

**The Randomized Posttest-Only Control Group Design, Using Matched Subjects**

| | | | |
|---|---|---|---|
| Treatment group | $M_r$ | $X$ | $O$ |
| Control group | $M_r$ | $C$ | $O$ |

**The Randomized Pretest-Posttest Control Group Design, Using Matched Subjects**

| | | | | |
|---|---|---|---|---|
| Treatment group | $M_r$ | $O$ | $X$ | $O$ |
| Control group | $M_r$ | $O$ | $C$ | $O$ |

The symbol $M_r$ refers to the fact that the members of each matched pair are randomly assigned to the experimental and control groups.

Although a pretest of the dependent variable is commonly used to provide scores on which to match, a measurement of any variable that shows a substantial relationship to the dependent variable is appropriate. Matching may be done in either or both of two ways: mechanically or statistically. Both require a score for each subject on *each* variable on which subjects are to be matched.

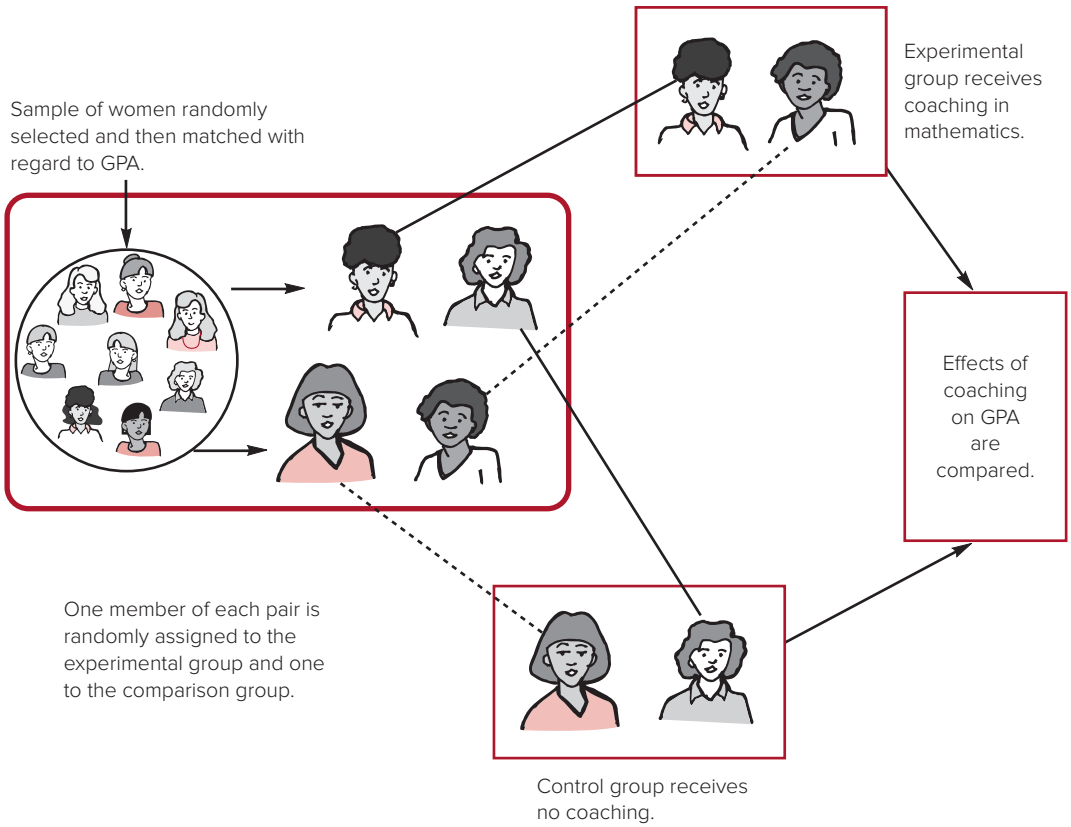



**Figure 13.7** *A Randomized Posttest-Only Control Group Design, Using Matched Subjects*

**Mechanical matching** is a process of pairing two persons whose scores on a particular variable are similar. Two girls, for example, whose mathematics aptitude scores and test anxiety scores are similar might be matched on those variables. After the matching is completed for the entire sample, a check should be made (through the use of frequency polygons) to ensure that the two groups are indeed equivalent on each matching variable. Unfortunately, two problems limit the usefulness of mechanical matching. First, it is very difficult to match on more than two or three variables—people just don't pair up on more than a few characteristics, making it necessary to have a very large initial sample to draw from. Second, in order to match, it is almost inevitable that some subjects must be eliminated from the study because no "matches" for them can be found. Samples then are no longer random even though they may have been before matching occurred.

As an example of a mechanical matching design with random assignment, suppose a researcher is interested in the effects of academic coaching on the grade-point averages (GPA) of low-achieving students in science classes. The researcher randomly selects a sample of 60 students from a population of 125 such students in a local elementary school and matches them by pairs on GPA, finding that she can match 40 of the 60. She then randomly assigns each subject in the resulting 20 pairs to either the experimental or the control group. Figure 13.7 presents a similar example.

**Statistical matching**,* on the other hand, does not necessitate a loss of subjects, nor does it limit the number of matching variables. Each subject is given a "predicted" score on the dependent variable, based on the correlation between the dependent variable and the variable (or variables) on which the subjects are being matched. The difference between the predicted and actual scores for each individual is then used to compare experimental and control groups.

*__Statistical equating__ is a more common term than its synonym, *statistical matching*. We believe the meaning for the beginning student is better conveyed by the term *matching*.

When a pretest is used as the matching variable, the difference between the predicted and actual score is called a **regressed gain score**. This score is preferable to the more straightforward **gain scores** (posttest minus pretest score for each individual) primarily because it is more reliable. We discuss a similar procedure under partial correlation in Chapter 15.

If mechanical matching is used, one member of each matched pair is randomly assigned to the experimental group, the other to the control group. If statistical matching is used, the sample is divided randomly at the outset, and the statistical adjustments are made after all data have been collected. Although some researchers advocate the use of statistical over mechanical matching, statistical matching is not infallible. Its major weakness is that it assumes that the relationship between the dependent variable and each predictor variable can be properly described by a straight line rather than a curved line. Whichever procedure is used, the researcher must (in this design) rely on random assignment to equate groups on all other variables related to the dependent variable.

## QUASI-EXPERIMENTAL DESIGNS

**Quasi-experimental designs** do not include the use of random assignment. Researchers who employ these designs rely instead on other techniques to control (or at least reduce) threats to internal validity. We shall describe some of these techniques as we discuss several quasi-experimental designs.

**The Matching-Only Design.** The **matching-only design** differs from random assignment with matching only in the fact that random assignment is not used. The researcher still matches the subjects in the experimental and control groups on certain variables, but he or she has no assurance that they are equivalent on others. Why? Because even though matched, subjects already are in intact groups. This is a serious limitation but often is unavoidable when random assignment is impossible—that is, when intact groups must be used. When several (say, 10 or more) groups are available for a method study and the groups can be randomly assigned to different treatments, this design offers an alternative to random assignment of subjects. After the groups have been randomly assigned to the different treatments, the individuals receiving one treatment are matched with individuals receiving the other treatments. The design shown in Figure 13.7 is still preferred, however.

It should be emphasized that matching (whether mechanical or statistical) is never a substitute for random assignment. Furthermore, the correlation between the matching variable(s) and the dependent variable should be fairly substantial. (We suggest at least .40.) Realize also that unless it is used in conjunction with random assignment, matching controls only for the variable(s) being matched. Diagrams of each of the matching-only control group designs follow:

**The Matching-Only Posttest-Only Control Group Design**

| | | | |
|---|---|---|---|
| Treatment group | *M* | *X* | *O* |
| Control group | *M* | *C* | *O* |

**The Matching-Only Pretest-Posttest Control Group Design**

| | | | | |
|---|---|---|---|---|
| Treatment group | *M* | *O* | *X* | *O* |
| Control group | *M* | *O* | *C* | *O* |

The *M* in this design means that the subjects in each group have been matched (on certain variables) but not randomly assigned to the groups.

**Counterbalanced Designs.** **Counterbalanced designs** represent another technique for equating experimental and comparison groups. In this design, each group is exposed to all treatments, however many there are, but in a different order. Any number of treatments may be involved. An example of a diagram for a counterbalanced design involving three treatments is as follows:

**A Three-Treatment Counterbalanced Design**

| | | | | | | |
|---|---|---|---|---|---|---|
| Group I | $X_1$ | *O* | $X_2$ | *O* | $X_3$ | *O* |
| Group II | $X_2$ | *O* | $X_3$ | *O* | $X_1$ | *O* |
| Group III | $X_3$ | *O* | $X_1$ | *O* | $X_2$ | *O* |

This arrangement involves three groups. Group I receives treatment 1 and is posttested, then receives treatment 2 and is posttested, and last receives treatment 3 and is posttested. Group II receives treatment 2 first, then treatment 3, and then treatment 1, being posttested after each treatment. Group III receives treatment 3 first, then treatment 1, followed by treatment 2, also being posttested after each treatment. The order in which the groups receive the treatments should be determined randomly.

How do researchers determine the effectiveness of the various treatments? Simply by comparing the average

| | Study 1 | | Study 2 | |
|---|---|---|---|---|
| | *Weeks 1-4* | *Weeks 5-8* | *Weeks 1-4* | *Weeks 5-8* |
| Group I | Method $X = 12$ | Method $Y = 8$ | Method $X = 10$ | Method $Y = 6$ |
| Group II | Method $Y = 8$ | Method $X = 12$ | Method $Y = 10$ | Method $X = 14$ |
| Overall Means: | Method $X = 12$; Method $Y = 8$ | | Method $X = 12$; Method $Y = 8$ | |

**Figure 13.8** *Results (Means) from a Study Using a Counterbalanced Design*

scores for all groups on the posttest for each treatment. In other words, the averaged posttest score for all groups for treatment 1 can be compared with the averaged posttest score for all groups for treatment 2, and so on, for however many treatments there are.

This design controls well for the subject characteristics threat to internal validity but is particularly vulnerable to multiple-treatment interference—that is, performance during a particular treatment may be affected by one or more of the previous treatments. Consequently, the results of any study in which the researcher has used a counterbalanced design must be examined carefully. Consider the two sets of hypothetical data shown in Figure 13.8.

The interpretation in study 1 is clear: Method *X* is superior for both groups regardless of sequence and to the same degree. The interpretation in study 2, however, is much more complex. Overall, method *X* appears superior, and by the same amount as in study 1. In both studies, the overall mean for *X* is 12, while for *Y* it is 8. In study 2, however, it appears that the difference between *X* and *Y* depends on previous exposure to the other method. Group I performed much worse on method *Y* when it was exposed to it following *X,* and group II performed much better on *X* when it was exposed to it after method *Y.* When either *X* or *Y* was given first in the sequence, there was no difference in performance. It is not clear that method *X* is superior in all conditions in study 2, whereas this is quite clear in study 1.

**Time-Series Designs.** The typical pre- and posttest designs examined up to now involve observations or measurements taken immediately before and after treatment. A **time-series design**, however, involves repeated measurements or observations over a period of time both before and after treatment. It is really an elaboration of the one-group pretest-posttest design presented in Figure 13.2. An extensive amount of data is collected on a single group. If the group scores essentially the same on the pretests and then considerably improves on the posttests, the researcher has more confidence that the treatment is causing the improvement than if just one pretest and one posttest were given. An example might be a teacher who gives a weekly test to her class for several weeks before giving them a new textbook to use, and then monitors how they score on a number of weekly tests after they have used the text. A diagram of the basic time-series design is as follows:

**A Basic Time-Series Design**

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

The threats to internal validity that endanger use of this design include history (something could happen between the last pretest and the first posttest), instrumentation (if, for some reason, the test being used is changed at any time during the study), and testing (due to a practice effect). The possibility of a pretest-treatment interaction is also increased with the use of several pretests.

The effectiveness of the treatment in a time-series design is basically determined by analyzing the pattern of test scores that results from the several tests. Figure 13.9 illustrates several possible outcome patterns that might result from the introduction of an experimental variable (*X*). The vertical line indicates the point at which the experimental treatment is introduced. In this figure, the change between time periods 5 and 6 gives the same kind of data that would be obtained using a one-group pretest-posttest design. The collection of additional data before and after the introduction of the treatment, however, shows how misleading a one-group pretest-posttest design can be. In (A), the improvement is shown to be no more than that which occurs from one data collection period to another—regardless of method. You will notice that performance does improve from time to time, but no trend or overall increase is

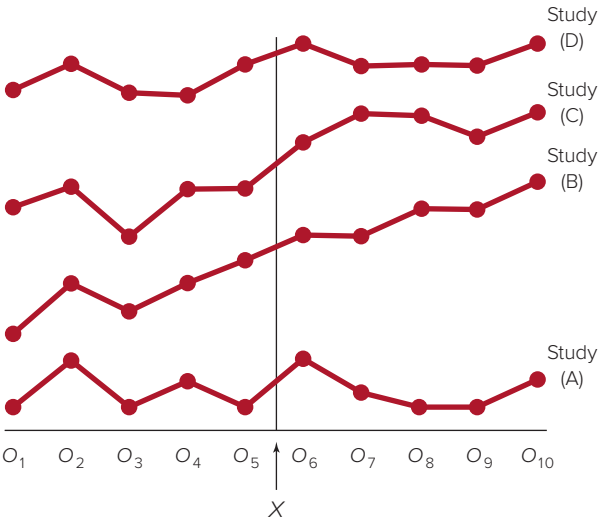



**Figure 13.9** *Possible Outcome Patterns in a Time-Series Design*

apparent. In (B), the gain from periods 5 to 6 appears to be part of a trend already apparent before the treatment was begun (quite possibly an example of maturation). In (D) the higher score in period 6 is only temporary, as performance soon approaches what it was before the treatment was introduced (suggesting an extraneous event of transient impact). Only in (C) do we have evidence of a consistent effect of the treatment.

The time-series design is a strong design, although it is vulnerable to history (an extraneous event could occur after period 5) and instrumentation (owing to the several test administrations at different points in time). The extensive amount of data collection required, in fact, is a likely reason why this design is infrequently used in educational research. In many studies, especially in schools, it simply is not feasible to give the same instrument eight to ten times. Even when it is possible, serious questions are raised concerning the validity of instrument interpretation with so many administrations. An exception is the use of unobtrusive devices that can be applied over many occasions, since interpretations based on them should remain valid.

## FACTORIAL DESIGNS

**Factorial designs** extend the number of relationships that may be examined in an experimental study. They are essentially modifications of either the posttest-only control group or pretest-posttest control group designs (with or without random assignment), which permit the investigation of additional independent variables. Another value of a factorial design is that it allows a researcher to study the **interaction** of an independent variable with one or more other variables, sometimes called *moderator variables.* **Moderator variables** may be either treatment variables or subject characteristic variables. A diagram of a factorial design is as follows:

**Factorial Design**

| | | | | | |
|---|---|---|---|---|---|
| Treatment | *R* | *O* | *X* | $Y_1$ | *O* |
| Control | *R* | *O* | *C* | $Y_1$ | *O* |
| Treatment | *R* | *O* | *X* | $Y_2$ | *O* |
| Control | *R* | *O* | *C* | $Y_2$ | *O* |

This design is a modification of the pretest-posttest control group design. It involves one treatment and one control group, and a moderator variable having two levels ($Y_1$ and $Y_2$). In this example, two groups would receive the treatment ($X$) and two would not ($C$). The groups receiving the treatment would differ on $Y$, however, as would the two groups not receiving the treatment. Because each variable, or factor, has two levels, the above design is called a 2 by 2 factorial design. This design can also be illustrated as follows:

**Alternative Illustration of the Preceding Example**

| | *X* | *C* |
|---|---|---|
| $Y_1$ | | |
| $Y_2$ | | |

A variation of this design uses two or more different treatment groups and no control groups. Consider the example we have used before of a researcher comparing the effectiveness of inquiry and lecture methods of instruction on achievement in history. The independent variable in this case (method of instruction) has two levels—inquiry ($X_1$) and lecture ($X_2$). Now imagine the researcher wants to see whether achievement is also influenced by class size. In that case, $Y_1$ might represent small classes and $Y_2$ might represent large classes.

As we suggest, it is possible using a factorial design to assess not only the separate effect of each independent variable but also their joint effect. In other words, the researcher is able to see how one of the variables might moderate the other (hence the reason for calling these variables *moderator variables*).

| | *Method* | |
|---|---|---|
| **Class size** | **Inquiry ($X_1$)** | **Lecture ($X_2$)** |
| Small ($Y_1$) | | |
| Large ($Y_2$) | | |

**Figure 13.10** *Using a Factorial Design to Study Effects of Method and Class Size on Achievement*

Let us continue with the example of the researcher who wished to investigate the effects of method of instruction and class size on achievement in history. Figure 13.10 illustrates how various combinations of these variables could be studied in a factorial design.

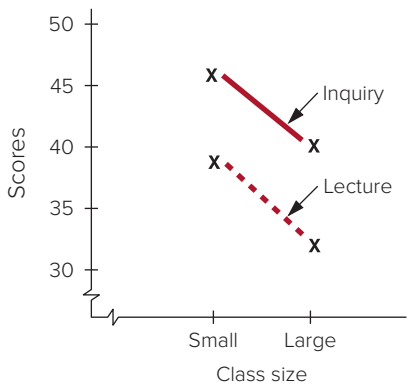
Factorial designs, therefore, are an efficient way to study several relationships with one set of data. Let us emphasize again, however, that their greatest virtue lies in the fact that they enable a researcher to study interactions between variables.

Figure 13.11, for example, illustrates two possible outcomes for the 2 by 2 factorial design shown in Figure 13.10. The scores for each group on the posttest (a 50-item quiz on American history) are shown in the boxes (usually called *cells*) corresponding to each combination of method and class size.

In study (*a*) in Figure 13.11, the inquiry method was shown to be superior in both small and large classes, and small classes were superior to large classes for both methods. Hence no interaction effect is present. In study (*b*), students did better in small than in large classes with both methods; however, students in small classes did better when they were taught by the inquiry method, but students in large classes did better when they were taught by the lecture method. Thus, even though students did better in small than in large classes in general, how well they did depended on the teaching method. As a result, the researcher cannot say that either method was always better; it depended on the size of the class in which students were taught. There was an interaction, in other

(*a*) No interaction between class size and method

| | Method | | |
|---|---|---|---|
| Class size | Inquiry ($X_1$) | Lecture ($X_2$) | Mean |
| Small ($Y_1$) | 46 | 38 | 42 |
| Large ($Y_2$) | 40 | 32 | 36 |
| Mean = | 43 | 35 | |




(*b*) Interaction between class size and method

| | Method | | |
|---|---|---|---|
| Class size | Inquiry ($X_1$) | Lecture ($X_2$) | Mean |
| Small ($Y_1$) | 48 | 42 | 45 |
| Large ($Y_2$) | 32 | 38 | 35 |
| Mean = | 40 | 40 | |

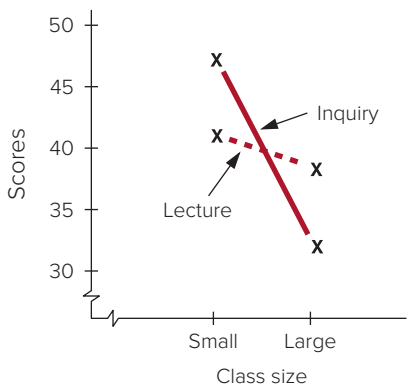



**Figure 13.11** *Illustration of Interaction and No Interaction in a 2 by 2 Factorial Design*

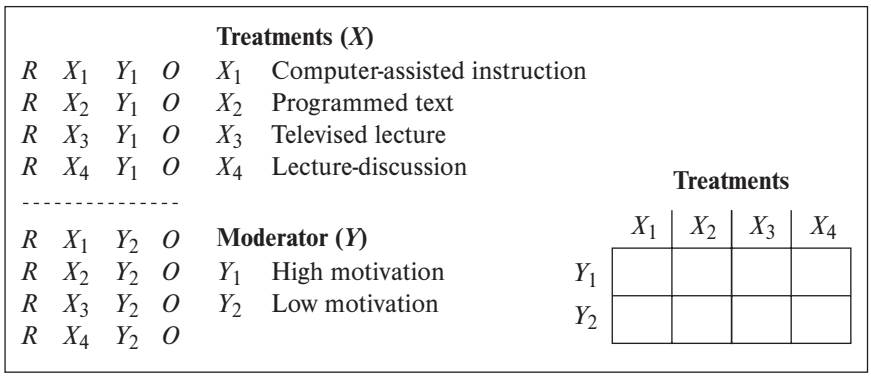



**Figure 13.12**
*Example of a 4 by 2 Factorial Design*

words, between class size and method, and this in turn affected achievement.

Suppose a factorial design was *not* used in study (*b*). If the researcher simply compared the effect of the two methods, without taking class size into account, he would have concluded that there was no difference in their effect on achievement (notice that the means of both groups = 40). The use of a factorial design enables us to see that the effectiveness of the method, in this case, depended on the size of the class in which it was used. It appears that an interaction existed between method and class size.

A factorial design involving four levels of the independent variable and using a modification of the posttest-only control group design was employed by Tuckman.[8] In this study, the independent variable was type of instruction, and the moderator was amount of motivation. It is a 4 by 2 factorial design (Figure 13.12). Many additional variations are also possible, such as 3 by 3, 4 by 3, and 3 by 2 by 3 designs. Factorial designs can be used to investigate more than two variables, although rarely are more than three variables studied in one design.

## Control of Threats to Internal Validity: A Summary

Table 13.1 presents our evaluation of the effectiveness of each of the preceding designs in controlling the threats to internal validity that we discussed in Chapter 9. You should remember that these assessments reflect our judgment; not all researchers would necessarily agree. We have assigned two pluses ( + +) to indicate a *strong* control (the threat is *unlikely* to occur); one plus ( +) to indicate *some* control (the threat *might* occur); a minus (−) to indicate a *weak* control (the threat is *likely* to occur); and a question mark (?) to those threats whose likelihood, owing to the nature of the study, we cannot determine.

You will notice that these designs are most effective in controlling the threats of subject characteristics, mortality, history, maturation, and regression. Note that mortality is controlled in several designs because any subject lost is lost to both the experimental and control methods, thus introducing no advantage to either. A location threat is a minor problem in the time-series design because the location where the treatment is administered is usually constant throughout the study; the same is true for data collector characteristics, although such characteristics may be a problem in other designs if different collectors are used for different methods. This is usually easy to control, however. Unfortunately, time-series designs do suffer from a strong likelihood of instrument decay and data collector bias, since data (by means of observations) must be collected over many trials, and the data collector can hardly be kept in the dark as to the intent of the study.

Unconscious bias on the part of data collectors is not controlled by any of these designs, nor is an implementation effect. Either implementers or data collectors can, unintentionally, distort the results of a study. The data collector should be kept ignorant as to who received which treatment, if this is feasible. It should be verified that the treatment is administered and the data collected as the researcher intended.

As you can see in Table 13.1, a testing threat may be present in many of the designs, although its magnitude depends on the nature and frequency of the instrumentation involved. It can occur only when subjects respond to an instrument on more than one occasion.

**TABLE 13.1** *Effectiveness of Experimental Designs in Controlling Threats to Internal Validity*

| | Threat | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Design** | **Subject Characteristics** | **Mortality** | **Location** | **Instrument Decay** | **Data Collector Characteristics** | **Data Collector Bias** | **Testing** | **History** | **Maturation** | **Attitude of Subjects** | **Regression** | **Implementation** |
| One-shot case study | – | – | – | (NA) | – | – | (NA) | – | – | – | – | – |
| One group pretest-posttest | – | ? | – | – | – | – | – | – | – | – | – | – |
| Static-group comparison | – | – | – | + | – | – | + | ? | + | – | – | – |
| Randomized posttest-only control group | ++ | + | – | + | – | – | ++ | + | ++ | – | ++ | – |
| Randomized pretest-posttest control group | ++ | + | – | + | – | – | + | + | ++ | – | ++ | – |
| Randomized Solomon four-group | ++ | ++ | – | + | – | – | ++ | + | ++ | – | ++ | – |
| Randomized posttest-only control group with matched subjects | ++ | + | – | + | – | – | ++ | + | ++ | – | ++ | – |
| Matching-only pretest-posttest control group | + | + | – | + | – | – | + | + | + | – | + | – |
| Counterbalanced | ++ | ++ | – | + | – | – | + | ++ | ++ | ++ | ++ | – |
| Time-series | ++ | – | + | – | + | + | – | – | + | – | ++ | – |
| Factorial with randomization | ++ | ++ | – | ++ | – | – | + | + | ++ | – | ++ | – |
| Factorial without randomization | ? | ? | – | ++ | – | – | + | + | + | – | ? | – |

*Key:* ( + +) = strong control, threat unlikely to occur; (+) = some control, threat may possibly occur; (–) = weak control, threat likely to occur; (?) = can't determine; (NA) = threat does not apply.

The attitudinal (or demoralization) effect is best controlled by the counterbalanced design since each subject receives both (or all) special treatments. In the remaining designs, it can be controlled by providing another "special" experience during the alternative treatment. Special mention should be made of the double-blind type of experiment. Such studies are common in medicine but hard to arrange in education. The key element is that neither the subjects nor the researcher knows the identity of those receiving each treatment. This is most easily accomplished in medical studies by means of a *placebo* (sometimes a sugar pill) that is indistinguishable from the actual medicine.

Regression is not likely to be a problem except in the one-group pretest-posttest design, since it should occur equally in experimental and control conditions if it

CONTROVERSIES IN RESEARCH

## Do Placebos Work?

The placebo effect—the expectation that some patients will show improvement if they are given any kind of treatment at all, even a sugar pill—has long been acknowledged by physicians and others involved in clinical trials. But does this effect really exist?

Two researchers in Denmark suggested it often does not. They reviewed 114 clinical trials in which patients were given real medicine, a placebo, or no treatment at all. Their report, published in the *New England Journal of Medicine* in May 2001, showed that "placebos offer no significant advantage over "no treatment" for dozens of conditions ranging from colds and seasickness to hypertension and Alzheimer's disease. (The exception was pain relief, which sugar pills seem to bring to about 15 percent of patients.)"* The researchers speculated that one explanation of the placebo effect may simply have been an unconscious desire by patients to please their doctors.

What do you think? Do some patients try (unconsciously) to please their doctors?

*Reported in *Time,* June 4, 2001, p. 65.

occurs at all. It could, however, possibly occur in a static-group pretest-posttest control group design, if there are large initial differences between the two groups.

## Evaluating the Likelihood of a Threat to Internal Validity in Experimental Studies

An important consideration in planning an experimental study or in evaluating the results of a reported study is the likelihood of threats to internal validity. As we have shown, a number of possible threats to internal validity may exist. The question that a researcher must ask is: How likely is it that any *particular* threat exists in *this* study?

To aid in assessing this likelihood, we suggest the following procedures.

**Step 1:** Ask: What specific factors either are known to affect the dependent variable or may logically be expected to affect this variable? (Note that researchers need *not* be concerned with factors unrelated to what they are studying.)

**Step 2:** Ask: What is the likelihood of the comparison groups differing on each of these factors? (A difference between groups cannot be explained away by a factor that is the same for all groups.)

**Step 3:** Evaluate the threats on the basis of how likely they are to have an effect, and plan to control for them. If a given threat cannot be controlled, acknowledge this.

The importance of step 2 is illustrated in Figure 13.13. In each diagram, the thermometers depict the performance of subjects receiving method A compared to those receiving method B. In diagram (*a*), subjects receiving method A performed higher on the posttest but *also* performed higher on the pretest; thus, the difference in pretest achievement accounts for the difference on the posttest. In diagram (*b*), subjects receiving method A performed higher on the posttest but did *not* perform higher on the pretest; thus, the posttest results *cannot* be explained by, or attributed to, different achievement levels prior to receiving the methods.

Let us consider an example to illustrate how these different steps might be employed. Suppose a researcher wishes to investigate the effects of two different teaching methods (e.g., lecture versus inquiry instruction) on critical thinking ability of students (as measured by scores on a critical thinking test). The researcher plans to compare two groups of eleventh-graders, one group being taught by an instructor who uses the lecture method, the other group being taught by an instructor who uses the inquiry method. Assume that intact classes will be used rather than random assignment to groups. Several of the threats to internal validity discussed in Chapter 9 are considered and evaluated using the steps just presented. We would argue that this is the kind of thinking researchers should engage in when planning a research project.

**Subject Characteristics.** Although many possible subject characteristics might affect critical thinking

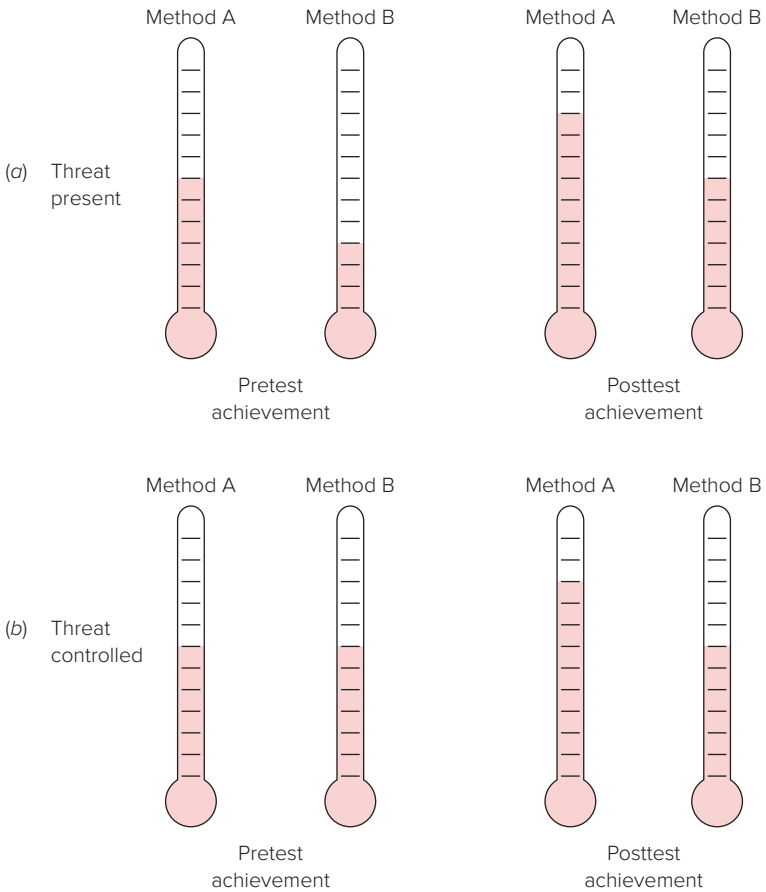



**Figure 13.13** *Guidelines for Handling Internal Validity in Comparison Group Studies*

ability, we identify only two here—(a) initial critical thinking ability and (b) gender.

1. **Critical thinking ability.** *Step 1:* Posttreatment critical thinking ability of students in the two groups is almost certainly related to initial critical thinking ability. *Step 2:* Groups may well differ unless randomly assigned or matched. *Step 3:* Likelihood of having an effect unless controlled: high.
2. **Gender.** *Step 1:* Posttreatment critical ability may be related to gender. *Step 2:* Groups may differ in proportions of each gender unless controlled by matching. *Step 3:* Likelihood of having an effect unless controlled: moderate.

**Mortality.** *Step 1:* Mortality is likely to affect posttreatment scores on any measure of critical thinking since those subjects who drop out or are otherwise lost would likely have lower scores. *Step 2:* Groups probably would not differ in numbers lost, but this should be verified. *Step 3:* Likelihood of having an effect unless controlled: moderate.

**Location.** *Step 1:* If location of implementation of treatment and/or of data collection differs for the two groups, this could affect posttreatment scores on the critical thinking test. Posttreatment scores would be expected to be affected by such resources as class size, availability of reading materials, films, and so forth. *Step 2:* This threat may differ for groups unless controlled for by standardizing locations for implementation and data collection. The classrooms using each method may differ systematically unless steps are taken to ensure that resources are comparable. *Step 3:* Likelihood of having an effect unless controlled: moderate to high.

MORE ABOUT RESEARCH

## Significant Findings in Experimental Research

In our opinion, some of the most important research in social psychology, with obvious implications for education, has been that on the effects of cooperative social interaction on negative attitudes, or the tendency of people to dislike others. A series of experimental studies begun in the 1940s led to the generalization that liking for group members, including those of different backgrounds and ethnicity, is increased by cooperative activities that lead to a successful outcome.* An application of this finding is the "jigsaw technique," which requires each member of a group to teach other members a section of material to be learned.† Experimental studies generally support the effectiveness of this procedure.

*Stephan, W. G. (1985). Intergroup relations. In G. Lindzey and E. Aronson (Eds.), *Handbook of social psychology.* New York: Random House.
†Aronson, E., Stephan, C., Sikes, J., Blaney, N., & Snapp, M. (1978). *The jigsaw classroom.* Beverly Hills: Sage.

**Instrumentation.**

1. **Instrument decay.** *Step 1:* Instrument decay may affect any outcome. *Step 2:* Instrument decay could differ for groups. This should not be a major problem, provided all instruments used are carefully examined and any alterations found are corrected. *Step 3:* Likelihood of having an effect unless controlled: low.
2. **Data collector characteristics.** *Step 1:* Data collector characteristics might affect scores on critical thinking test. *Step 2:* This threat might differ for groups unless controlled by using the same data collector(s) for all groups. *Step 3:* Likelihood of having an effect unless controlled: moderate.
3. **Data collector bias.** *Step 1:* Bias could certainly affect scores on critical thinking test. *Step 2:* This threat might differ for groups unless controlled by training implementers in administration of the instrument and/or keeping them ignorant as to which treatment group is being tested. *Step 3:* Likelihood of having an effect unless controlled: high.

**Testing.** *Step 1:* Pretesting, if used, might well affect posttest scores on critical thinking test. *Step 2:* Presumably the pretest would affect both groups equally, however, and would not be likely to interact with method, since instructors using each method are teaching critical thinking skills. *Step 3:* Likelihood of having an effect unless controlled: low.

**History.** *Step 1:* Extraneous events that might affect critical thinking skills are difficult to conjecture, but they might include such things as a special TV series on thinking, attendance at a district workshop on critical thinking by some students, or participation in certain extracurricular activities (e.g., debates) that occur during the course of the study. *Step 2:* In most cases, these events would likely affect both groups equally and hence are not likely to constitute a threat. Such events should be noted and their impact on each group assessed to the degree possible. *Step 3:* Likelihood of having an effect unless controlled: low.

**Maturation.** *Step 1:* Maturation could affect outcome scores since critical thinking is presumably related to individual growth. *Step 2:* Presuming that the instructors teach each method over the same time period, maturation should not be a threat. *Step 3:* Likelihood of having an effect unless controlled: low.

**Attitude of Subjects.** *Step 1:* Subjects' attitudes could affect posttest scores. *Step 2:* If the members of either group perceive that they are receiving any sort of "special attention," this could be a threat. The extent to which either treatment is "novel" should be evaluated. *Step 3:* Likelihood of having an effect unless controlled: low to moderate.

**Regression.** *Step 1:* Regression is unlikely to affect posttest scores unless subjects are selected on the basis of extreme scores. *Step 2:* This threat is unlikely to affect groups differently, although it could do so. *Step 3:* Likelihood of having an effect unless controlled: low.

**Implementation.** *Step 1:* Instructor characteristics are likely to affect posttreatment scores. *Step 2:*

Because different instructors teach the methods, they may well differ. This could be controlled by having several instructors for each method, by having each instructor teach both methods, or by monitoring instruction. *Step 3:* Likelihood of having an effect unless controlled: high.

The trick, then, to identifying threats to internal validity is, first, to think of different variables (conditions, subject characteristics, and so on) that might affect the outcome variable of the study and, second, to decide, based on evidence and/or experience, whether these things would affect the comparison groups differently. If so, the influence of these factors may provide an alternative explanation for the results. If this seems likely, a threat to internal validity of the study may indeed be present and needs to be minimized or eliminated. It should then be discussed in the final report on the research project.

## Control of Experimental Treatments

The designs discussed in this chapter are all intended to improve the internal validity of an experimental study. As you have seen, each has its advantages and disadvantages, and each provides a way of handling some threats but not others.

Another issue, however, cuts across all designs. While it has been touched on in earlier sections, particularly in connection with location and implementation threats, it deserves more attention than it customarily receives. The issue is that of researcher control over the experimental treatment(s). Of course, an essential requirement of a well-conducted experiment is that researchers have control over the treatment—that is, they control the what, who, when, and how of it. A clear example of researcher control is the testing of a new drug; clearly, the drug is the treatment and the researcher can control who administers it, under what conditions, when it is given, to whom, and how much. Unfortunately, researchers seldom have this degree of control in educational research.

In the ideal situation, a researcher can specify precisely the ingredients of the treatment; in actual practice, many treatments or methods are too complex to describe precisely. Consider the example we have previously given of a study comparing the effectiveness of inquiry and lecture methods of instruction. What, exactly, is the individual who implements each method to do? Researchers may differ greatly in their answers to this question. Ambiguity in specifying exactly what the implementer of the treatment is to do leads to major problems in implementation. How are researchers to train teachers to implement the methods involved in a study if they can't specify the essential characteristics of those methods? Even supposing that adequate specification can be achieved and training methods developed, how can researchers be sure the methods are implemented *correctly?* These problems must be faced by any researcher using any of the designs we have discussed.

A consideration of this issue frequently leads to consideration (and assessment) of possible trade-offs. The greatest control is likely to occur when the researcher is the one implementing the treatment; this, however, also provides the greatest opportunity for an implementation threat to occur. The more the researcher diffuses implementation by adding other implementers in the interest of reducing threats, however, the more he or she risks distortion or dilution of the treatment. The extreme case is the use of existing treatment groups—that is, groups located by the researcher that already are receiving certain treatments. Most authors refer to these as causal-comparative or *ex post facto* studies (see Chapter 16), and do not consider them to fall under the category of experimental research. In such studies, the researcher must locate groups receiving the specified treatment(s) and then use a matching-only design or, if sufficient lead time exists before implementation of the treatment, a time-series design. We are not persuaded that such studies, if treatments are carefully identified, are necessarily inferior with respect to cause-effect conclusions compared with studies in which treatments are assigned to teachers (or others) by the researcher. Both are equally open to most of the threats we have discussed. The existing groups are more susceptible to subject characteristics, location, and regression threats than true experiments, but not necessarily more so than quasi-experiments. One would expect fewer problems with an attitudinal effect, since existing practice is not altered. Greater history and maturation threats exist because the researcher would have less control. Implementation is difficult to assess. Teachers who are already implementing a new method may be enthusiastic if they initially chose the method, but they also may be better teachers. On the other hand, teachers assigned to a method that is new to them may be either enthusiastic or resentful. We conclude that both types of study are legitimate.