CHAPTER 5

Introduction to Experimental Research

Preview & Chapter Objectives

The middle four chapters of this text, Chapters 5 through 8, concern the design of experiments. The first half of Chapter 5 outlines the essential features of an experiment—varying some factor of interest (the independent variable), controlling all other factors (extraneous variables), and measuring the outcome (dependent variables). In the second part of this chapter, you will learn how the validity of a study can be affected by how well it is designed. When you finish this chapter, you should be able to:

- Define a manipulated independent variable and identify examples that are situational, task, and instructional variables.
- Distinguish between experimental and control groups.
- Describe John Stuart Mill's rules of inductive logic and apply them to the concepts of experimental and control groups.

- Recognize the presence of confounding variables in an experiment and understand why confounding creates serious problems for interpreting the results of an experiment.
- Distinguish independent from dependent variables, given a brief description of any experiment.
- Distinguish between independent variables that are manipulated variables or subject variables, and understand the interpretation problems that accompany the use of subject variables.
- Recognize the factors that can reduce the statistical conclusion validity of an experiment.
- Describe how construct validity applies to the design of an experiment.
- Describe the various ways in which an experiment's external validity can be reduced.
- Describe and be able to recognize the various threats to an experiment's internal validity.
- Recognize that external validity might not be important for all research but that internal validity is essential.
- Understand the ethical guidelines for running a "subject pool."

When Robert Sessions Woodworth finally published *Experimental Psychology* in 1938, the book's contents were already well known among psychologists. As early as 1909, Woodworth was giving his Columbia University students copies of a mimeographed handout called "Problems and Methods in Psychology," and a companion handout called "Laboratory Manual: Experiments in Memory, etc." appeared in 1912. By 1920, the manuscript filled 285 pages and was called "A Textbook of Experimental Psychology." After a 1932 revision, still in mimeograph form, the book finally was published in 1938. By then Woodworth's students were using it to teach their own students, and it was so widely known that the publisher's announcement of its publication said simply, "The Bible Is Out" (Winston, 1990).

The so-called Columbia bible was encyclopedic, with more than 823 pages of text and another 36 pages of references. After an introductory chapter, it was organized into 29 different research topics such as "memory," "maze learning," "reaction time," "association," "hearing," "the perception of color," and "thinking." Students wading through the text would learn about the methods used in each content area, and they would also learn virtually everything there was to know in 1938 about each topic.

The impact of the Columbia bible on the teaching of experimental psychology has been incalculable. Indeed, the teaching of experimental psychology today, and to some degree the structure of the book you're now reading, are largely cast in the mold set by Woodworth. In particular, he took the term "experiment," until then loosely defined as virtually any type of empirical research, and gave it the definition it has today. In particular, he contrasted experimental with correlational research, a distinction now taken for granted.

The defining feature of the experimental method was the manipulation of what Woodworth called an "independent variable," which would affect what he called

the "dependent variable." In his words, the experimenter "holds all the conditions constant except for one factor which is his 'experimental factor' or his 'independent variable.' The observed effect is the 'dependent variable' which in a psychological experiment is some characteristic of behavior or reported experience" (Woodworth, 1938, p. 2). Although the terms were not invented by Woodworth, he was the first to use them as they are used today.

While the experimental method manipulates independent variables, the correlational method, according to Woodworth, "[m]easures two or more characteristics of the same individuals [and] computes the correlation of these characteristics. This method... has no 'independent variable' but treats all the measured variables alike" (Woodworth, 1938, p. 3). You will learn more about correlational research in Chapter 9. In this and the next three chapters, however, the focus will be on the experimental method, the researcher's most powerful tool for identifying cause-and-effect relationships.

Essential Features of Experimental Research

Since Woodworth's time, psychologists have thought of an **experiment** as a systematic research study in which the investigator directly varies some variable (or variables), holds all other factors constant, and observes the results of the systematic variation. The factors under the control of the experimenter are called independent variables, the variables being held constant are referred to as extraneous variables, and the behaviors measured are the dependent variables. Before we examine these concepts more closely, however, you should read Box 5.1, which describes the logical foundations of the experimental method in a set of rules proposed by the British philosopher John Stuart Mill in 1843.

Box 5.1

ORIGINS—John Stuart Mill and the Rules of Inductive Logic

John Stuart Mill (1805–1873) was England's preeminent nineteenth-century philosopher. Although he was known primarily as a political philosopher, much of his work has direct relevance for psychology. For example, his book on *The Subjection of Women* (1869) argued forcefully and well ahead of its time that women had abilities equal to those of men and ought to be treated equally with men. Of importance for our focus on methodology, in 1843 he published

A System of Logic, Ratiocinative and Inductive, Being a Connected View of the Principles of Evidence, and the Methods of Scientific Investigation (in those days, they liked to pack all they could into a title!). In his *Logic*, Mill argued for the creation of a science of psychology (he called it "ethology") on the grounds that, while it might not reach the level of precision of physics, it could do just as well as other disciplines that were considered scientific at the time (meteorology was the example he used). He also laid out a set of methods that form the logical basis for what you will learn in this chapter and in the chapter on correlation. The methods were those of 'Agreement" and "Difference" (relevant for this chapter), and of "Concomitant Variation" (relevant for correlation—see Chapter 9, pp. 311–313).

Taken together, the methods of Agreement and Difference enable us to conclude, with a high degree of confidence, that some outcome, Y, was caused by some factor, X. The Method of Agreement states that if X is regularly followed by Y, then X is *sufficient* for Y to occur, and could be a cause of Y. That is, "if X, then Y." The Method of Difference states that if Y does not occur when X does not occur, then X is *necessary* for Y to occur—"if not X, then not Y." Taken together (what Mill called the "Joint Method"), the methods of Agreement and Difference provide the necessary and sufficient conditions (i.e., the immediate cause) for the production of Y.

To make this more concrete, suppose we are trying to determine if watching violent TV causes a child to be aggressive. "Watching violent TV" is X, and "aggression" is Y. If we can determine that every time a child watches violent TV (X), the result is some act of aggression (Y), then we have satisfied the method of Agreement, and we can say that watching violent TV is enough (sufficient) to produce aggression. If the child watches violent TV, then aggression occurs ("If X, then Y"). If we can also show that whenever violent TV is not watched (not X), the child is not aggressive (not Y), then we can say that watching violent TV is necessary in order for aggression to occur. If the child does not watch violent TV, aggression does not occur ("If not X, then not Y").

It is important to note that in the real world of research, the conditions described in these methods are never met fully. That is, it will be impossible to identify and measure the outcome of every instance of every child watching TV. Rather, the best one can do is to observe systematically as many instances as possible, under controlled conditions, and then draw conclusions with a certain amount of confidence. That is precisely what research psychologists do and, as you recall from the Chapter 1 discussion of scientific thinking, the reason why researchers regard all knowledge based on science to be tentative, pending additional research. As findings are replicated, confidence in them increases.

As you work through this chapter, especially at the point where you learn about studies with experimental and control groups, you will see that an experimental group (e.g., some children shown violent TV shows) accomplishes Mill's Method of Agreement, while a control group (e.g., other children not shown violent films) accomplishes the Method of Difference. Studies with both experimental and control groups meet the conditions of Mill's Joint Method.

Establishing Independent Variables

Any experiment can be described as a study investigating the effect of X on Y. The "X" is Woodworth's **independent variable**: it is the factor of interest to the experimenter, the one that is being studied to see if it will influence behavior. It is sometimes called a "manipulated" factor because the experimenter has complete control over it and is creating the situations that research participants will encounter in the study. As you will see, the concept of an independent variable can also be stretched to cover what are called nonmanipulated or subject variables, but, for now, let us consider only those independent variables that are under the experimenter's total control.

Independent variables must have a minimum of two *levels*. That is, at the very least, an experiment involves a comparison between two situations (or *conditions*). For example, suppose a researcher is interested in the effects of different dosages of marijuana on reaction time. In such a study, there have to be at least two different dosage levels in order to make a comparison. This study would be described as an experiment with "amount of marijuana" as the independent variable and "dosage 1" and "dosage 2" as the two levels of the independent variable. You could also say that the study has two conditions in it—the two dosage levels. Of course, independent variables can have more than two levels. In fact, there are distinct advantages to adding levels beyond the minimum of two, as you will see in Chapter 7 on experimental design.

Experimental research can be either basic or applied in its goals, and it can be conducted either in the laboratory or in the field (refer back to Chapter 3, pp. 78–83 for an elaboration of these distinctions). Experiments that take place in the field are sometimes called **field experiments**. The term **field research** is a broader term for any empirical research outside of the laboratory, including both experimental studies and studies using nonexperimental methods.

Varieties of Independent Variables

The range of factors that can be used as independent variables is limited only by the creative thinking of the researcher. However, independent variables that are manipulated in a study tend to fall into three somewhat overlapping categories: situational variables, task variables, and instructional variables.

Situational variables refer to different features in the environment that participants might encounter. For example, in a helping behavior study, the researcher interested in studying the effect of the number of bystanders on the chances of help being offered might create a situation in which participants encounter a person in need of help. Sometimes the participant is alone with the person needing aid; at other times the participant and the victim are accompanied by a group of either three or six bystanders. In this case, the situational independent variable would be the number of potential helpers on the scene besides the participant, and the levels would be zero, three, and six bystanders.

Sometimes experimenters vary the type of task performed by participants. One way to manipulate **task variables** is to give groups of participants different kinds of problems to solve. For instance, research on the psychology of reasoning often

involves giving people different kinds of logical problems to determine the kinds of errors people tend to make. Similarly, mazes can differ in the degree of complexity, different types of illusions could be presented in a perception study, and so on.

Instructional variables are manipulated by asking different groups to perform a particular task in different ways. For example, children in a memory task who are all shown the same list of words might be given different instructions about how to memorize the list. Some might be told to form visual images of the words, others might be told to form associations between adjacent pairs of words, and still others might be told simply to repeat each word three times as it is presented.

Of course, it is possible to combine several types of independent variables in a single study. A study of the effects of crowding, task difficulty, and motivation on problem-solving ability could have participants placed in either a large or a small room, thereby manipulating crowding through the situational variable of room size. Some participants in each type of room could be given difficult crossword puzzles to solve and others less difficult ones; this illustrates a task variable. Finally, an instructional variable could manipulate motivation by telling participants that they will earn either $1 or $5 for completing the puzzles.

Control Groups

In some experiments, the independent variable is whether or not some treatment is administered. The levels of the independent variable in this case are essentially 1 and 0; some get the treatment and others don't. In a study of the effects of TV violence on children's aggressive behavior, for instance, some children might be shown a violent TV program, while others don't get to see it, or see a nonviolent TV show. The term **experimental group** is used as a label for the first situation, in which the treatment is present. Those in the second type of condition, in which treatment is withheld, are said to be in the **control group**. Ideally, the participants in a control group are identical to those in the experimental group in all ways except that the control group participants do not get the experimental treatment. As you recall from Box 5.1, the conditions of the experimental group satisfy Mill's Method of Agreement (if violent TV, then aggression) and the control group can satisfy the Method of Difference (if no violent TV, then no aggression). Thus, a simple experiment with an experimental and a control group is an example of what Mill called the "Joint Method." In essence, the control group provides a baseline measure against which the experimental group's behavior can be compared. Think of it this way: control group = comparison group.

Please don't think that control groups are necessary in all research, however. It is indeed important to *control* extraneous variables, as you are about to learn, but control *groups* occur in research only when it is important to have a comparison with a baseline level of performance. For example, suppose you were interested in the construct "sense of direction," and wanted to know whether a specific training program would help people avoid getting lost in new environments. In that study, a reasonable comparison would be between a training group and a control group without training. On the other hand, if your empirical question concerns gender differences in sense of direction, the comparison will be between a group of males and a group of females—neither would be considered a control group. You will

learn about several specialized types of control groups in Chapter 7, the first of two chapters dealing with experimental design.

Controlling Extraneous Variables

The second feature of the experimental method is that the researcher tries to control what are called **extraneous variables**. These are any variables that are not of interest to the researcher but which might influence the behavior being studied if they are not controlled properly. As long as these are held constant, they present no danger to the study. If they are not adequately controlled, however, they might influence the behavior being measured in some systematic way. The result is called confounding. A **confound** is any uncontrolled extraneous variable that "covaries" with the independent variable and could provide an alternative explanation of the results. That is, a confounding variable changes at the same time that an independent variable changes (i.e., they "covary") and, consequently, its effect cannot be separated from the effect of the independent variable. Hence, when a study has a confound, the results could be due to the effects of *either* the confounding variable or the independent variable, or some combination of the two, and there is no way to decide among these alternatives.

To illustrate some obvious confounding, consider a verbal learning experiment in which a researcher wants to show that students who try to learn a large amount of material all at once don't do as well as those who spread their learning over several sessions. That is, massed practice (cramming?) is predicted to be inferior to distributed practice. Three groups of students are selected, and each group is given the same five chapters in a general psychology text to learn. Participants in the first group are given 3 hours on Monday to study the material. Participants in the second group are given 3 hours on Monday and 3 hours on Tuesday, and those in the final group get 3 hours each on Monday, Tuesday, and Wednesday. On Friday, all the groups are tested on the material (see Table 5.1 for the design). The results show that Group 3 scores the highest, followed by Group 2. Group 1 does not do well at all, and the researcher concludes that distributed practice is superior to massed practice. Do you agree with this conclusion?

You probably don't, because there are two serious confounds in this study, both easy to spot. The participants certainly differ in how their practice is distributed (1, 2, or 3 days), but they *also* differ in how much total practice they get (3, 6, or

TABLE 5.1 *Confounding in a Hypothetical Distribution of Practice Experiment*

	Monday	Tuesday	Wednesday	Thursday	Friday
Group 1	3	—	—	—	Exam
Group 2	3	3	—	—	Exam
Group 3	3	3	3	—	Exam

Note: The 3 in each equals the number of hours spent studying five chapters of a general psychology text.

TABLE 5.2 *Identifying Confounds*

Levels of IV Distribution of Practice	EV 1 Study Hours	EV 2 Retention Interval	DV Retention Test Performance
1 day	3 hours	3 days	Lousy
2 days	6 hours	2 days	Average
3 days	9 hours	1 day	Great

IV = independendt variable.
EV = extraneous variable.
DV = dependent variable.

9 hours). This is a perfect example of a confound—it is impossible to tell if the results are due to one factor (distribution of practice) or the other (total practice hours); the two factors covary perfectly. The way to describe this situation is to say "distribution of practice is confounded with total study hours." The second confound is perhaps less obvious but is equally problematic. It concerns the retention interval. The test is on Friday for everyone, but different amounts of time have elapsed between study and test for each group. Perhaps Group 3 did the best because they studied the material most recently and forgot the least amount. In this experiment, distribution of practice is confounded both with total study hours and with retention interval. Each confound by itself could account for the results, and the factors may also have interacted with each other in some way to provide yet another interpretation.

Look at Table 5.2, which gives you a convenient way to identify confounds. In the first column are the levels of the independent variable and in the final column are the results. The middle columns are extraneous variables that should be held constant through the use of appropriate controls. If they are not kept constant, then confounding exists. As you can see for the distributed practice example, the results could be explained by the variation in any of the first three columns, either individually or in some combination. To correct the confound problem in this case, you need to ensure that the middle two columns are constant instead of variable.

A problem that students sometimes have with understanding confounds is that they tend to use the term whenever they spot something in a study that might not be right. For example, suppose the distribution of practice study included the statement that only females were used in the study. Some students reading the description might think there's a confound here—gender. What they really mean is they believe both males and females ought to be in the study and that might indeed be the case, but gender is *not* a confound in this example. Gender would be a confound only if males were used just in one condition and only females were used in one other condition. Then any group differences in the results could be due to the independent variable or to gender. So be careful. A confound is a serious flaw in a study, but not all design flaws are confounds.

In the Applications exercises at the end of the chapter you will be identifying confounds. You might find the task easier if you fit the problems into the Table 5.2 format. Take a minute and redesign the distributed practice study. How would you eliminate the confounding from these extraneous variables?

Learning to be aware of potential confounding factors and building appropriate ways to control for them is one of the scientific thinking skills that is most difficult to develop. Not all confounds are as obvious as the massed/distributed practice example. We'll encounter the problem often in the remaining chapters and address it again shortly in the context of a discussion of what is called the internal validity of a study.

Measuring Dependent Variables

The third part of any experiment is measuring some behavior that is presumably being influenced by the independent variable. The term **dependent variable** is used to describe those behaviors that are the measured outcomes of experiments. If, as mentioned earlier, an experiment can be described as the effect of X on Y and "X" is the independent variable, then "Y" is the dependent variable. In a study of the effects of TV violence on children's aggressiveness, the dependent variable would be some measure of aggressiveness. In the distribution of practice study, it would be a measure of exam performance.

The credibility of any experiment and its chances of discovering anything of value depend partly on the decisions made about what behaviors to measure as dependent variables. We've already seen that empirical questions cannot be answered unless the terms are defined with some precision. You might take a minute and review the section on operational definitions in Chapter 3 (pp. 85–86). When an experiment is designed, one key component concerns the operational definitions for the behaviors to be measured as dependent variables. Unless the behaviors are defined precisely, replication is impossible.

Deciding on dependent variables can be tricky. A useful guide is to know the prior research and use already-established dependent measures, those that have been shown to be reliable and valid. Sometimes you have to develop a new measure, however, and when you do, a brief pilot study might help you avoid two major problems that can occur with poorly chosen dependent variables—ceiling and floor effects. A **ceiling effect** occurs when the average scores for the different groups in the study are so high that no difference can be determined. This happens when your dependent measure is so easy that everyone gets a high score. Conversely, a **floor effect** happens when all the scores are extremely low because the task is too difficult for everyone, once again producing a failure to find any differences between groups.

One final point about variables. It is important to realize that a particular construct could be an independent, an extraneous, *or* a dependent variable, depending on the research problem at hand. An experiment might manipulate a particular construct as an independent variable, try to control it as an extraneous factor, or measure it as a dependent variable. Consider the construct of anxiety, for instance. It could be a manipulated independent variable by telling participants that they will be experiencing shocks that will be either moderate or painful when they make errors on a simulated driving task. Anxiety could also be a factor that needs to be held constant in some experiments. For instance, if you wanted to evaluate the effects of a public speaking workshop on the ability of students to deliver a brief speech, you wouldn't want to videotape the students in one group without taping those in the other group as well. If everyone is taped, then the level of anxiety created by that factor (taping)

is held constant for everyone. Finally, anxiety could be a dependent variable in a study of the effects of different types of exams (e.g., multiple choice vs. essay) on the perceived test anxiety of students during final exam week. Some physiological measures of anxiety might be used in this case. Anxiety could also be considered a personality characteristic, with some people having more of it than others; this last possibility leads to the next topic.

✓ Self Test 5.1

1. In a study of the effects of problem difficulty (easy or hard) and reward size ($1 or $5 for each solution) on an anagram problem-solving task, what are the independent and dependent variables?
2. What are extraneous variables and what happens if they are not controlled properly?
3. Explain how frustration could be an independent, extraneous, or dependent variable, depending on the study.

Manipulated versus Subject Variables

Up to this point, the term independent variable has meant some factor manipulated directly by the researcher. An experiment compares one condition created by and under the control of the experimenter with another. However, in many studies, comparisons are also made between groups of people who differ from each other in ways other than those designed by the researcher. These comparisons are made between factors that are referred to variously as ex post facto variables, natural group variables, nonmanipulated variables, or **subject variables,** which is the term I will use. They refer to already existing characteristics of the individuals participating in the study, such as gender, age, socioeconomic class, cultural group, intelligence, physical or psychiatric disorder, and any personality attribute you can name. When using subject variables in a study, the researcher cannot manipulate them directly but must *select* people for the different conditions of the experiment by virtue of the characteristics they already have.

To illustrate the differences between manipulated and subject variables, consider a hypothetical study of the effects of anxiety on maze learning in humans. You could *manipulate* anxiety directly by creating a situation in which one group is made anxious (told they'll be performing in front of a large audience perhaps), while a second group is not (no audience). In that study, any person who volunteers could potentially wind up in one group or the other. To do the study using a *subject* variable, on the other hand, you would select two groups differing in their characteristic levels of anxiety and ask each to try the maze. The first group would be those who were anxious types of people (as determined ahead of time by a personality test for anxiety proneness). The second group would include more relaxed types of people. Notice

the major difference between this situation and one involving a manipulated variable. With anxiety as a subject variable, volunteers coming into the study cannot be placed into either of the conditions (anxious-all-the-time-Fred cannot be put into the low-anxiety group), but must be in one group or the other, depending on attributes they *already* possess prior to entering the study.

Some researchers, true to Woodworth's original use of the term, prefer to reserve the term independent variable for those variables directly manipulated by the experimenter. Others are willing to include subject variables as examples of a particular type of independent variable on the grounds that the experimenter has some degree of control over them by virtue of the decisions involved in selecting them in the first place. I take this latter position and will use the term independent variable in the broader sense. However, whether this term is used broadly (manipulated + subject) or narrowly (manipulated only) is not important, providing you understand the difference between a manipulated and a nonmanipulated or subject variable.

Research Example 4—Using Subject Variables

One common type of research using subject variables examines differences from one culture to another. Ji, Peng, and Nisbett (2000) provide a nice example. In a series of studies involving various cognitive tasks, they looked at the implications of the differences between those raised in Asian cultures and those raised in Western cultures. In general, they pointed out that Asians, especially those from China, Korea, and Japan, have a "relatively holistic orientation, emphasizing relationships and connectedness" (p. 943) among objects, rather than on the individual properties of the objects themselves. Those from Western cultures, especially those deriving from the Greek "analytic" tradition, are "prone to focus more exclusively on the object, searching for those attributes of the object that would help explain and control its behavior" (p. 943).

This cultural difference led Ji et al. (2000) to make several predictions, including one that produced a study with two separate subject variables—culture and gender. They chose a cognitive task that has a long history, the rod and frame test (RFT). While sitting in a darkened room, participants in an RFT study see an illuminated square frame projected on a screen in front of them, along with a separate illuminated straight line (rod) inside the frame. The frame can be oriented to various angles by the experimenter and the participant's task is to move a device that changes the orientation of the rod. The goal is to make the rod perfectly vertical, regardless of the frame's orientation. The classic finding (Witkin & Goodenough, 1977) is that some people (field independent) are quite able to bring the rod into a true vertical position, disregarding the distraction of the frame, while others (field dependent) adjust the rod with reference to the frame and not with reference to true vertical. Can you guess the hypothesis? The researchers predicted that those from Asian cultures would be more likely to be field dependent than those from Western cultures. They also hypothesized greater field dependence for females, a prediction based on a typical finding in RFT studies. So, in terms of the concepts introduced in Chapter 3 (pp. 102–103), part of this study (gender) involved replication and part (culture) involved extension.

Because the undergraduate population of the University of Michigan (where the study was conducted) includes a large number of East Asians, Ji et al. (2000) were

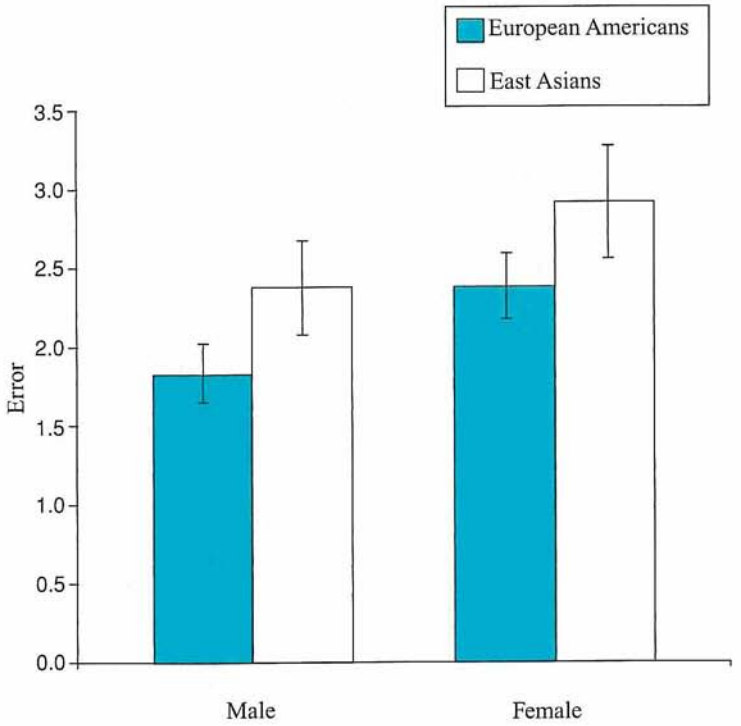

FIGURE 5.1 Gender and cultural differences in the rod and frame test, from Ji, Peng, and Nisbett's (2000) cross-cultural study. Note the vertical lines at the top of each bar; these are called "error bars," and they reflect variability around the mean (see pp. xx).

able to complete their study using students enrolled in general psychology classes there (in a few pages you'll be learning about "subject pools"). They compared 56 European Americans with 42 East Asians (most from China, Korea, and Japan) who had been living in the United States for an average of about 2.5 years. Students in the two cultural groups were matched in terms of SAT math scores, and there were about an equal number of males and females in each group.

As you can see from Figure 5.1, the results supported both hypotheses. The finding about females being more field dependent than males was replicated, and the difference occurred in both cultures. In addition, the main finding was the consistent difference between the cultures—those from East Asian cultures were more field dependent than the European Americans. As Ji et al. (2000) described the outcome, the relative field independence of the Americans reflected their tendency to be "more attentive to the object and its relation to the self than to the field" (p. 951), while the field dependent Asians tended to be "more attentive to the field and to the relationship between the object and the field" (p. 952). One statistical point worth noting relates to the concept of an outlier, introduced in Chapter 4 (p. 137). Each subject did the RFT task 16 times and, on average, 1.2 of their scores were omitted from the analysis because they were significantly beyond the normal

range of scores. Their operational definition of outlier was somewhat technical, but related to the distance from the interquartile range, another concept you recall from Chapter 4 (p. 138).

Only a study using *manipulated* independent variables can be called an experiment in the strictest sense of the term; it is sometimes called a "true" experiment (which sounds a bit pretentious and carries the unfortunate implication that other studies are "false"). Studies using independent variables that are *subject* variables are occasionally called ex post facto studies or quasi experiments ("quasi" meaning "to some degree" here).[1] Sometimes (often, actually) studies will include both manipulated and subject independent variables. Being aware of the presence of subject variables is important because they affect the kinds of conclusions that can be drawn from the study's results.

Drawing Conclusions When Using Subject Variables

Put a little asterisk next to this section—it is extremely important. Recall from Chapter 1 that one of the goals of research in psychology is to discover explanations for behavior. That is, we wish to know what caused some behavior to occur. Simply put, with manipulated variables, conclusions about the causes of behavior can be made; with subject variables, they cannot. The reason has to do with the amount of control held by the experimenter in each case.

With manipulated variables, the experiment can meet the criteria listed in Chapter 1 for demonstrating causality. The independent variable precedes the dependent variable, covaries with it, and, assuming that no confounds are present, can be considered the most reasonable explanation for the results. In other words, if you vary some factor and successfully hold all else constant, the results can be attributed *only* to the factor varied. In a confound-free experimental study with two groups, these groups will be essentially equal to each other (i.e., any differences will be random ones) in all ways except for the manipulated factor.

When using subject variables, however, the experimenter can also vary some factor (i.e., select participants having certain characteristics) but cannot hold all else constant. Selecting participants who are high or low on some definition of anxiety proneness does not guarantee that the two groups will be equivalent in other ways. In fact, they might be different from each other in several ways (in self-confidence, perhaps) that could influence the outcome of the study. When a difference between the groups occurs in this type of study, we cannot say that the differences were *caused* by the subject variable. In terms of the conditions for causality, while we can say that the independent variable precedes the dependent variable and covaries with it, we cannot eliminate alternative explanations for the relationship because certain extraneous factors cannot be controlled. When subject variables are present, all we can say is that the groups performed differently on the dependent measure.

---

[1]The term quasi-experimental design is actually a broader designation referring to any type of design in which participants cannot be randomly assigned to the groups being studied (Cook & Campbell, 1979). These designs are often found in applied research and are elaborated in Chapter 10.

An example from social psychology might help to clarify the distinction. Suppose you were interested in altruistic behavior and wanted to see how it was affected by the construct of "self-esteem." The study could be done in two ways. First, you could manipulate self-esteem directly by first giving participants a personality test. By providing different kinds of false feedback about the results of the test, both positive and negative, self-esteem could be raised or lowered temporarily. The participants could then be asked to do some volunteer work to see if those feeling good about themselves would be more likely to help.[2] A second way to do this study is to give participants a reliable and valid personality test for level of self-esteem and select those who score in the upper 25% and lower 25% on the measure as the participants for the two groups. Self-esteem in this case is a subject variable—half of the participants will be low self-esteem types, while the other half will be high self-esteem types. As in the first study, these two groups of people could be asked about volunteering.

In the first study, differences in volunteering can be traced *directly* to the self-esteem manipulation. If all other factors are properly controlled, the temporary feeling of increased or decreased self-esteem is the *only* thing that could have produced the differences in helping. In the second study, however, you cannot say that high self-esteem is the direct cause of the helping behavior; what you can say is that people with high self-esteem are more likely to help than those with low self-esteem. All you can do is to speculate about the reasons why this might be true because these participants may differ from each other in other ways unknown to you. For instance, high self-esteem types of people might have had prior experience in volunteering, and this experience might have had the joint effect of raising or strengthening their self-esteem and increasing the chances that they will volunteer in the future. Or they might have greater expertise in the specific volunteering tasks (e.g., public speaking skills). As you will see in Chapter 9, this difficulty in interpreting research with subject variables is exactly the same problem encountered when trying to draw conclusions from correlational research.

Returning for a moment to the Ji, Peng, and Nisbett (2000) study, which featured the subject variables of culture and gender, the authors were careful to avoid drawing conclusions about causality. The word "cause" never appears in their article, and the descriptions of results are always in the form "this group scored higher than this other group."

Before moving on to the discussion of the validity of experimental research, read Box 5.2. It identifies the variables in a classic study that you probably recall from your general psychology course—one of the so-called Bobo experiments that first investigated imitative aggression. Working through the example will help you apply your knowledge of independent, extraneous, and dependent variables, and will allow you to see how manipulated and subject variables are often encountered in the same study.

---

[2]Manipulating self-esteem raises ethical questions that were considered in a study by Sullivan and Deiker (1973). See Chapter 2, p. 58.

Box 5.2

CLASSIC STUDIES—Bobo Dolls and Aggression

Ask any student who has just completed a course in child, social, or personality psychology (perhaps even general psychology) to tell you about the Bobo doll studies. The response will be immediate recognition and a brief description along the lines of "Oh, yes, the studies showing that children will punch out an inflated doll if they see an adult doing it." A description of one of these studies is a good way to clarify further the differences between independent, extraneous, and dependent variables. The study was published by Albert Bandura and his colleagues in 1963 and is entitled "Imitation of Film-Mediated Aggressive Models" (Bandura, Ross, & Ross, 1963).

Establishing Independent Variables

The study included both manipulated and subject variables. The major manipulated variable was the type of experience that preceded the opportunity for aggression. There were four levels, including three experimental groups and a control group.

Experimental group 1: real-life aggression (children directly observed an adult model aggressing against the Bobo doll)

Experimental group 2: human film aggression (children observed a film of an adult model aggressing against Bobo)

Experimental group 3: cartoon film aggression (children observed a cartoon of "Herman the Cat" aggressing against a cartoon Bobo)

Control group: no exposure to aggressive models

The nonmanipulated independent variable (subject variable) was gender. Male and female students from the Stanford University Nursery School (mean age = 52 months) were the participants in the study. (Actually, there was also another manipulated variable; participants in groups 1 and 2 were exposed to either a same-gender or opposite-gender model.) The basic procedure of the experiment was to expose the children to some type of aggressive model (or not, for the control group), and then put them into a room full of toys (including Bobo), thereby giving them the opportunity to be aggressive themselves.

Controlling Extraneous Variables

Several possible confounds were avoided. First, in groups 1 and 2, the adults aggressed against a *5-foot* Bobo doll. When given a chance to pummel Bobo themselves, the children were put into a room with a *3-foot* Bobo doll. This kept the size relationship between person and doll approximately constant. Second, participants in all four groups were mildly frustrated before being given a chance to aggress. They were

allowed to play for a few minutes with some very attractive toys and then were told by the experimenter that the toys were special and were being reserved for some other children. Thus, for *all* of the children, there was an approximately equivalent increase in their degree of emotional arousal just prior to the time when they were given the opportunity to be aggressive. Any differences in aggressiveness could be attributed to the imitative effects and not to any emotional differences between the groups.

Measuring Dependent Variables

Several different measures of aggression were used in this study. Aggressive responses were categorized as imitative, partially imitative, or nonimitative, depending on how closely they matched the model's behavior. For example, the operational definition of imitative aggressive behaviors included striking the doll with a wooden mallet, punching it in the nose, and kicking it. Partially imitative behaviors included hitting something else with the mallet and sitting on the doll but not hitting it. Nonimitative aggression included shooting darts from an available dart gun at targets other than Bobo and acting aggressively toward other objects in the room.

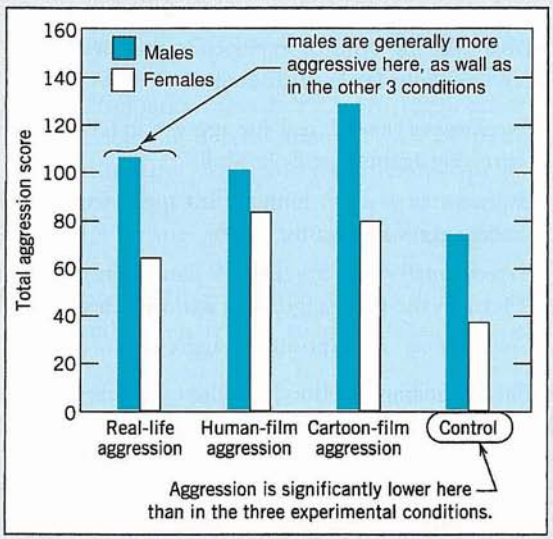

FIGURE 5.2 Data from Bandura, Ross, and Ross's Bobo study (1963) of the effects of imitation on aggression.

Briefly, the results of the study were that children in groups 1, 2, and 3 showed significantly more aggression than those in the control group, but the same amount of overall aggression occurred regardless of the type of modeling. Also, boys were more aggressive than girls in all conditions; some gender differences also occurred in the form of the aggression: girls "were more inclined than boys to sit on the Bobo doll but [unlike the boys] refrained from punching it" (Bandura et al., 1963, p. 9). Figure 5.2 summarizes the results.

The Validity of Experimental Research

Chapter 4 introduced the concept of validity in the context of measurement. The term also applies to experiments as a whole. Just as a measure is valid if it measures what it is supposed to measure, psychological research is said to be valid if it provides the understanding about behavior that it is supposed to provide. This section of the chapter introduces four different types of validity, following the scheme outlined by Cook and Campbell (1979) for research in field settings but applicable to any research in psychology. The four types of validity are statistical conclusion validity, construct validity (again), external validity, and internal validity.

Statistical Conclusion Validity

The previous chapter introduced you to the use of statistics in psychology. In particular, you learned about measurement scales, the distinction between descriptive and inferential statistics, and the basics of hypothesis testing. **Statistical conclusion validity** concerns the extent to which the researcher uses statistics properly and draws the appropriate conclusions from the statistical analysis.

The statistical validity of a study can be reduced in several ways. First, researchers might do the wrong analysis or violate some of the assumptions required for performing a particular analysis. For instance, the data for a study might be measured using an ordinal scale, thereby requiring the use of a particular type of statistical procedure. The researcher, however, mistakenly uses an analysis that is appropriate only for interval or ratio data. Second, the researcher might selectively report some analyses that came out as predicted but might not report others (guess which ones?), a practice that borders on fraud (see Chapter 2, pp. 68–70). The third example of a factor that reduces the statistical validity of a study concerns the reliability of the measures used. If the dependent measures are not reliable, there will be a great deal of error variability, which reduces the chances of finding a significant effect. If a true effect exits (i.e., H_0 should be rejected), but low reliability results in a failure to find that effect, the outcome would be a Type II error.

The careful researcher decides on the statistical analysis at the same time that the experimental design is being planned. In fact, no experiment should ever be designed without giving thought to how the data will be analyzed.

Construct Validity

The previous chapter described construct validity in the context of measuring psychological constructs: it refers to whether a test truly measures some construct (e.g., self-efficacy, connectedness to nature). In experimental research, **construct validity** has a related meaning: it refers to the adequacy of the operational definitions for *both* the independent and the dependent variables used in the study. In a study of the effects of TV violence on children's aggression, questions about construct validity could be (a) whether the programs chosen by the experimenter are the best choices to contrast violent with nonviolent television programming, and (b) whether the

operational definitions and measures of aggression used are the best ones that could be chosen. If the study used violent cartoon characters (e.g., Elmer Fudd shooting at Bugs Bunny) compared to nonviolent characters (e.g., Winnie the Pooh), someone might argue that children's aggressive behavior is unaffected by fantasy; hence, a more *valid* manipulation of the independent variable, called "level of filmed violence," would involve showing children realistic films of people that varied in the amount of violence portrayed.

Similarly, someone might criticize the appropriateness of a measure of aggression used in a particular study. This, in fact, has been a problem in research on aggression. For rather obvious ethical reasons, you cannot design a study that results in subjects punching each other's lights out. Instead, aggression has been defined operationally in a variety of ways, some of which might seem to you to be more valid (e.g., angered participants believing they are delivering shocks to another person) than others (e.g., horn honking by frustrated drivers). As was true for the discussion of construct validity in the previous chapter when the emphasis was on measurement, the validity of the choices about exactly how to define independent and dependent variables develops over time as accumulated research fits into a coherent pattern.

External Validity

Experimental psychologists have been occasionally criticized for knowing a great deal about college sophomores and white rats and very little about anything else. This is, in essence, a criticism of **external validity,** the degree to which research findings generalize beyond the specific context of the experiment being conducted. For research to achieve the highest degree of external validity, it is argued, its results should generalize in three ways—to other populations, to other environments, and to other times.

Other Populations

The comment about rats and sophomores fits here. As we have seen in Chapter 2, part of the debate over the appropriateness of animal research has to do with how well this research provides explanations that are relevant for human behavior. Concerning sophomores, recall that Milgram deliberately avoided using college students, and selected adults from the general population as subjects for his obedience studies. The same cannot be said of most social psychologists, however. A survey by Sears (1986) of research in social psychology found that 75% of the research published in 1980 used undergraduates as participants. When Sears repeated the survey for research published in 1985, the number was 74%. And it is not just social psychologists whose studies feature a high percentage of college students—since it began publication in 1992, 86% of the empirical articles in the *Journal of Consumer Psychology* have used college student samples (Jaffe, 2005). Sears argued that the characteristics of college students as a population could very well bias the general conclusions about social phenomena. Compared to the general population, for instance, college students are more able cognitively, more self-centered, more susceptible to social influence, and more likely to change their attitudes on issues. To the extent that research investigates issues related to those features, results from students might not generalize

to other groups, according to Sears. He suggested that researchers expand their databases and replicate important findings on a variety of populations. However, he also pointed out that many research areas (e.g., perception, cognition) produce outcomes relatively unaffected by the special characteristics of college students, and there is no question that students exist in large numbers and are readily available. Some special ethical considerations apply when using this group, as outlined in Box 5.3.

Box 5.3

ETHICS—Recruiting Participants: Everyone's in the Pool

Most research psychologists are employed by colleges and universities and consequently find themselves surrounded by an available supply of participants for their research. Because students may not readily volunteer to participate in research, most university psychology departments establish what is called the **subject pool** or the **participant pool**. The term refers to a group of students, typically those enrolled in introductory psychology classes, who are asked to participate in research as part of a course requirement. If you are a student at a large university, you have probably had the experience of "volunteering" for two or three experiments in order to avoid losing points or acquiring a grade of Incomplete for the course. At a large university, if 800 students take general psychology each semester and each student signs up for three studies, that makes 2,400 participants available to researchers.

Subject pools are convenient for researchers, and they are defended on the grounds that research participation is part of the educational process (Kimmel, 1996). Ideally, students can acquire deeper insights into the research process by being in the middle of experiments and learning something about the psychological phenomena being investigated. To maintain the "voluntary" nature, students are given the opportunity to complete the requirement with alternatives other than direct research participation. Problems exist, however. Critics argue that the pools are not really voluntary, that alternative activities (e.g., writing papers) are often so onerous and time-consuming that students are effectively compelled to sign up for the research, and that the research experience is more likely to be tedious and meaningless than educational (Korn, 1988). Some research supports such concern. A study by Sieber and Saks (1989) found evidence that 89% of 366 departments surveyed had pools that failed to meet at least one of the APA's recommendations (below).

Despite the potential for abuse, many psychology departments try to make the research experience educational for students. For example, during debriefing for a memory experiment, the participant/student could be told how the study relates to the information in Chapter X of the text being used in the introductory course.

Many departments also include creative alternative activities. These include having nonparticipating students (a) observe ongoing studies and record their observations, (b) participate in some community volunteer work, or (c) attend a research presentation by a visiting scholar and write a brief summary of it (Kimmel, 1996; McCord, 1991). Some studies have shown that students generally find research participation valuable, especially if researchers make an explicit attempt to tie the participation to the education occurring in the general psychology course (e.g., Landrum & Chastain, 1999; Leak, 1981).

The APA (1982, pp. 47–48) has provided some explicit guidelines about recruiting students as research participants, the main points being these:

- ✓ Students should be aware of the requirement before signing up for the course.
- ✓ Students should get a thorough description of the requirement on the first day of class, including a clear description of alternative activities if they opt not to serve as research subjects.
- ✓ Alternative activities must equal research participation in time and effort and, like participation, must have some educational value.
- ✓ All proposals for research using subject pools must have prior IRB approval;
- ✓ Special effort must be made to treat students courteously.
- ✓ There must be a clear and simple procedure for students to complain about mistreatment without their course grade being affected.
- ✓ All other aspects of the APA ethics code must be rigorously followed.
- ✓ The psychology department must have a mechanism in place to provide periodic review of pool policies.

The "college sophomore problem" is only one example of the concern over generalizing to other groups. Another has to do with gender. Some of psychology's most famous research has been limited by using only males (or, less frequently, only females), but drawing conclusions as if they apply to everyone. Perhaps the best-known example is Lawrence Kohlberg's research on children's moral development. Kohlberg (1964) asked adolescent boys (aged 10–16) to read and respond to brief accounts of various moral dilemmas. On the basis of the boys' responses, Kohlberg developed a six-stage theory of moral development that has become a fixture in developmental psychology texts. At the most advanced stage, the person acts according to a set of universal principles based on preserving justice and individual rights.

Kohlberg's theory has been criticized on external validity grounds. For example, Gilligan (1982) argued that Kohlberg's model overlooks important gender differences in thinking patterns and in how moral decisions are made. Males may come to place

the highest value on individual rights, but females tend to value the preservation of individual relationships. Hence, females responding to some of Kohlberg's moral dilemmas might not seem to be as morally advanced as males, but this is due to a biasing of the entire model because Kohlberg sampled only males, according to Gilligan.

Research psychologists also are careful about generalizing results from one culture to another. For example, "individualist" cultures are said to emphasize the unique person over the group, and personal responsibility and initiative are valued. On the other hand, the group is more important than the individual in "collectivist" cultures (Triandis, 1995). Research conclusions based on just one culture might not be universally applicable. To take just one example, most children in the United States are taught to place great value on personal achievement. In Japan, on the other hand, children learn that if they stand out from the crowd, they might diminish the value of others in the group; individual achievement is not as valuable. One study found that personal achievement was associated with positive emotions for American students, but with *negative* emotions for Japanese students (Kitayama, Markus, Matsumoto, & Norasakkunkit, 1997). To conclude that feeling good about individual achievement is a universal human trait would be a mistake. Does this mean that all research in psychology should make cross-cultural comparisons? No. It just means that conclusions sometimes need to be drawn cautiously, and with reference only to the group studied in the research project.

Other Environments

Besides generalizing to other types of individuals, externally valid results are applicable to other stimulus settings. This problem is the basis for the occasional criticism of laboratory research mentioned in Chapter 3—it is sometimes said to be artificial and too far removed from real life. Recall from the discussion of basic and applied research (pp. 78–80) that the laboratory researcher's response to criticisms about artificiality is to use Aronson's concept of experimental reality. The important thing is that people are involved in the study; mundane reality is secondary. In addition, laboratory researchers argue that some research is designed purely for theory testing and, as such, whether the results apply to real-life settings is less relevant than whether the results provide a good test of the theory (Mook, 1983).

Nonetheless, important developments in many areas of psychology have resulted from attempts to study psychological phenomena in real-life settings. A good example concerns the history of research on human memory. For much of the twentieth century, memory research occurred largely in the laboratory, where countless college sophomores memorized seemingly endless lists of words, nonsense syllables, strings of digits, and so on. The research created a comprehensive body of knowledge about basic memory processes that has value for the development of theories about memory and cognition, but whether principles discovered in the lab generalized to real-life memory situations was not clear. Change occurred in the 1970s, led by Cornell's Ulric Neisser. In *Cognition and Reality* (1976), he argued that the laboratory tradition in cognitive psychology, while producing important results, nonetheless had failed to yield enough useful information about information processing in real-world contexts. He called for more research concerning what he referred to as **ecological validity**—research with relevance for the everyday cognitive

activities of people trying to adapt to their environment. Experimental psychologists, Neisser urged, "must make a greater effort to understand cognition as it occurs in the ordinary environment and in the context of natural purposeful activity. This would not mean an end to laboratory experiments, but a commitment to the study of variables that are ecologically important rather than those that are easily manageable" (p. 7).

Neisser's call to arms was embraced by many (but not all, of course) cognitive researchers, and the 1980s and 1990s saw increased study of such topics as eyewitness memory (e.g., Loftus, 1979) and the long-term recall of subjects learned in school, such as Spanish (e.g., Bahrick, 1984). Neisser himself completed an interesting analysis of the memory of John Dean (Neisser, 1981), the White House chief counsel who blew the whistle on President Richard Nixon's attempted cover-up of illegal activities in the Watergate scandal of the early 1970s. Dean's testimony before Congress precipitated the scandal and led to Nixon's resignation. Dean's 245-page account was so detailed that some reporters referred to him as a human tape recorder. As you might know, it was later revealed that the Oval Office meetings described by Dean were also tape-recorded by the somewhat paranoid White House. Comparing the tapes with Dean's testimony gave Neisser a perfect opportunity to evaluate Dean's supposedly photographic memory, which turned out to be not so photographic after all—he recalled the general topics of the meetings reasonably well but missed a lot of the details and was often confused about sequences of events. The important point for external validity is that Neisser's study is a good illustration of how our knowledge of memory can be enriched by studying phenomena outside of the normal laboratory environment.

Other Times

The third way in which external validity is sometimes questioned has to do with the longevity of results. Some of the most famous experiments in the history of psychology are the conformity studies done by Solomon Asch in the 1950s (e.g., Asch, 1956). These experiments were completed during a historical period when conservative values were dominant in the United States, the "red menace" of the Soviet Union was a force to be concerned about, and conformity and obedience to authority were valued in American society. In that context, Asch found that college students were remarkably susceptible to conformity pressures. Would the same be true today? Would the factors that Asch found to influence conformity (e.g., group consensus) operate in the same way now? In general, research concerned with more fundamental processes (e.g., cognition) stands the test of time better than research involving social factors that may be embedded in some historical context.

A Note of Caution

Although external validity has value under many circumstances, it is important to point out that it is not always a major concern of research, and some (e.g., Mook, 1983) have even criticized the use of the term, because it carries the implication that research low in external "validity" is therefore "invalid." Yet there are many examples of research, completed in the laboratory under so-called artificial conditions, that have great value for the understanding of human behavior. Consider research on "false memory," for example (Roediger & McDermott, 1995). The

typical laboratory strategy is to give people a list of words to memorize, including a number of words from the same category—"sleep," for instance. The list might include the words dream, bed, pillow, nap, and so on, but not the broader term sleep. When recalling the list, many people recall the word sleep and they are often confident that the word was on the list when they are given a recognition test. That is, a laboratory paradigm exists demonstrating that people can sometimes remember something with confidence that they did not experience. The phenomenon has relevance for eyewitness memory (jurors pay more attention to confident eyewitnesses), but the procedure is far removed from an eyewitness context. It might be judged by some to be low in external validity. Yet there is important research going on that explores the theoretical basis for false memory, determining, for instance, the limits of the phenomenon and exactly how it occurs (e.g., Goodwin, Meissner, & Ericsson, 2001). That research will eventually produce a body of knowledge that comprehensively explains the false memory phenomenon.

In summary, the external validity of some research finding increases as it applies to other people, places, and times. But must researchers design a study that includes many different groups of people, takes place in several settings, including "realistic" ones, and gets repeated every decade? Of course not. External validity is not determined by an individual research project—it develops over time as research is replicated in various contexts—and as we have just seen, it is not always a relevant concern for research that is theory-based. Indeed, for the researcher designing a study, considerations of external validity pale compared to the importance of our next topic.

Internal Validity

The final type of experimental validity described by Cook and Campbell (1979) is called **internal validity**—the degree to which an experiment is methodologically sound and confound-free. In an internally valid study, the researcher feels confident that the results, as measured by the dependent variable, are directly associated with the independent variable and are not the result of some other, uncontrolled factor. In a study with confounding factors, as we've already seen in the massed/distributed practice example, the results will be uninterpretable. The outcome could be the result of the independent variable, the confounding variable(s), or some combination of both, and there is no clear way to decide between the different interpretations. Such a study would be quite low in internal validity.

✓ Self Test 5.2

1. Explain how "anxiety" could be both a manipulated variable and a subject variable.
2. In the famous "Bobo doll" study, what were the manipulated and the subject variables?
3. What is the basic difference between internal and external validity?
4. The study on the memory of John Dean was used to illustrate which form of validity?

Threats to Internal Validity

Any uncontrolled extraneous factor (i.e., confound) can reduce a study's internal validity, but there are a number of problems that require special notice (Cook & Campbell, 1979). These "threats" to internal validity are especially dangerous when control groups are absent, a problem that sometimes occurs in program evaluation research (Chapter 10). Many of these threats occur in studies that extend over a period of time during which several measures are taken. For example, participants might receive a pretest, an experimental treatment of some kind, and then a posttest. Ideally, the treatment should produce some positive effect that can be assessed by observing changes from the pretest to the posttest. A second general type of threat occurs when comparisons are made between groups that are said to be "nonequivalent." These so-called subject selection problems can interact with the other threats.

Pre-Post Studies

Do students learn general psychology better if the course is self-paced and computerized? If a college institutes a program to reduce test anxiety, can it be shown that it works? If you train people in various mnemonic strategies, will it improve their memories? These are all empirical questions that ask whether people will change as the result of some experience (a course, a program, memory training). To judge whether change occurred, one typical procedure is to evaluate people prior to the experience with what is known as a **pretest**. Then, after the experience, some **posttest** measure is taken. The ideal outcome for the examples I've just described is that, on the posttest, people (a) know general psychology better than they did at the outset, (b) are less anxious in test taking than they were before, or (c) show improvement in their memory. The typical research design compares experimental and control groups, with the latter not experiencing the treatment:

Experimental::	pretest	*treatment*	posttest
Control::	pretest		posttest

In the absence of a control group, there are several threats to the interval validity of research using pretests. Suppose we are trying to evaluate the effectiveness of a college's program to help students who suffer from test anxiety (i.e., they have decent study skills and seem to know the material, but they are so anxious during exams that they don't perform well on them). During orientation, first-year students fill out several questionnaires, including one that serves as a pretest for test anxiety. Let's assume that the scores can range from 20 to 100, with higher scores indicating greater anxiety. Incoming students who score high are asked to participate in the college's test anxiety program, which includes relaxation training, study skills training, and other techniques. Three months later they are assessed again for test anxiety, and the results look like this:

pretest	*treatment*	posttest
90		70

Thus, the average pretest score of those selected for the program is 90, and the average posttest score is 70. Assuming that the difference is statistically significant, what would you conclude? Did the treatment program work? Was the change due to the treatment, or could other factors have been involved? I hope you can see that there are several ways of interpreting this outcome. Read on.

History and Maturation

Sometimes an event occurs between pre- and posttesting that produces large changes unrelated to the treatment program; when this happens, the study is confounded by the threat of **history**. For example, suppose the college in the above example decided that grades are counterproductive to learning and that all courses would henceforth be graded on a pass/fail basis. Furthermore, suppose this decision came after the pretest for test anxiety and in the middle of the treatment program for reducing anxiety. The posttest might show a huge drop in anxiety, but this result could very likely be due to the historical event of the college's change in grading policy rather than to the program. Wouldn't you be a little more relaxed about this research methods course if grades weren't an issue?

In a similar fashion, the program for test anxiety involves first-year students at the very start of their college careers, so pre-post changes could also be the result of a general **maturation** of these students as they become accustomed to college life. As you probably recall, the first semester of college was a time of real change in your life. Maturation is always a concern whenever a study extends over some period of time.

Notice that if a control group is used, the experimenter can account for the effects of both history and maturation. These effects can be ruled out and the test anxiety program deemed effective if these results occurred:

Experimental::	**pretest**	**treatment**	**posttest**
	90		70
Control::	**pretest**		**posttest**
	90		90

On the other hand, either history or maturation or both would have to be considered as explanations for the changes in the experimental group if the control group scores also dropped to 70 on the posttest.

Regression

To regress is to go back, in this case in the direction of a mean score. Hence, the phenomenon I'm about to describe is sometimes called **regression to the mean**. In essence it refers to the fact that if score 1 is an extreme score, then score 2 will be closer to whatever the mean for the larger set of scores is. This is because, for

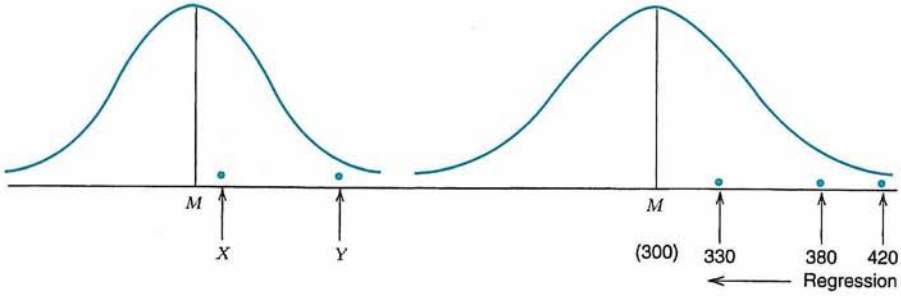

FIGURE 5.3 Regression to the mean.

a large set of scores, most will cluster around the mean and only a few will be far removed from the mean (i.e., extreme scores). Imagine you are selecting some score randomly from the normal distribution in Figure 5.3. Most of the scores center on the mean; so, if you make a random selection, you'll most likely choose a score near the mean (X on the left-hand side of Figure 5.3). However, suppose you just happen to select one that is far removed from the mean (i.e., an extreme score—Y). If you then choose again, are you most likely to pick

a. the exact same extreme score again?
b. a score even more extreme than the first one?
c. a score less extreme (i.e., closer to the mean) than the first one?

My guess is that you've chosen alternative "c," which means that you understand the basic concept of regression to the mean. To take a more concrete example (refer to the right-hand side of Figure 5.3), suppose you know that on the average (based on several hundred throws), Ted can throw a baseball 300 feet. Then he throws one 380 feet. If you were betting on his *next* throw, where would you put your money?

a. 380 feet
b. 420 feet
c. 330 feet

Again, I imagine you've chosen "c," further convincing yourself that you get the idea of the regression phenomenon. But what does this have to do with our pretest-posttest study?

In a number of pre-post studies, people are selected for some treatment because they've made an *extreme* score on the pretest. Thus, in the test anxiety study, participants were picked because on the pretest they scored very high for anxiety. On the posttest, their anxiety scores might improve (i.e., they will be lower than on the pretest), but the improvement could be a regression effect rather than the result of the memory improvement program. Once again, a control group of equivalent high-anxiety participants would enable the researcher to spot a possible regression effect. For instance, the following outcome would suggest that some regression might

be involved,[3] but the program nonetheless had an effect over and above regression. Can you see why this is so?

Experimental::	**pretest**	*treatment*	**posttest**
	90		70
Control::	**pretest**		**posttest**
	90		80

Regression effects can cause a number of problems, and were probably the culprit in some early studies that erroneously questioned the effectiveness of the well-known Head Start program. That particular example will be taken up in Chapter 10 as an example of some of the problems involved in assessing large-scale, federally supported programs.

Testing and Instrumentation

Testing is considered to be a threat to internal validity when the mere fact of taking the pretest has an effect on posttest scores. There could be a practice effect of repeated testing, or some aspects of the pretest could sensitize participants to something about the program. For example, if the treatment program is a self-paced, computerized general psychology course, the pretest would be some test of knowledge. Participants might be sensitized by the pretest to topics about which they seem to know nothing; they could then pay more attention to those topics during the course and do better on the posttest as a result.

Instrumentation is a problem when there are changes in the measurement instrument from pretest to posttest. In the self-paced general psychology course mentioned earlier, the pretest and posttest wouldn't be the same but would presumably be equivalent in level of difficulty. However, if the posttest happened to be easier, it would produce improvement that was more apparent than real. Instrumentation is sometimes a problem when the measurement tool involves observations. Those doing the observing might get better at it with practice, making the posttest instrument essentially different (more accurate in this case) from the pretest instrument.

Like the problems of history, maturation, and regression, the possible confounds of testing and instrumentation can be accounted for by including a control group. The only exception is that in the case of pretest sensitization, the experimental group might have a slight advantage over the control group on the posttest because the knowledge gained from the pretest might enable the experimental participants to focus on specific weaknesses during the treatment phase, whereas the control participants would not have that opportunity.

Participant Problems

Threats to internal validity can also arise from concerns over the individuals participating in the study. In particular, Cook and Campbell (1979) identified two problems.

---

[3]Notice that the sentence reads, "might be involved," not "must be involved." This is because it is also possible that the control group's change from 90 to 80 could be due to one of the other threats. Regression would be suspected if these other threats could not be ruled out.

Subject Selection Effects

One of the defining features of an experimental study with a manipulated independent variable is that participants in the different conditions are equivalent to each other except for the independent variable. In the next chapter you will learn how these equivalent groups are formed through random assignment and matching. If groups are not equivalent, then **subject selection** effects might occur. For example, suppose two sections of a general psychology course are being offered and a researcher wants to compare a traditional lecture course with the one combining lecture and discussion groups. School policy (a) prevents the researcher from randomly assigning students to the two courses, and (b) requires full disclosure of the nature of the courses. Thus, students can sign up for either section. You can see the difficulty here. If students in the lecture plus discussion course outperform students in the straight lecture course, what caused the difference? Was it the nature of the course (the discussion element) or was it something about the students who *chose* that course? Maybe they were more articulate (hence, interested in discussion) than those in the straight lecture course. In short, there is a confound due to the selection of subjects for the two groups being compared.

Selection effects can also interact with other threats to internal validity. For example, in a study with two groups, some historical event might affect one group but not the other. This would be referred to as a history x selection confound (read as "history by selection"). Similarly, two groups might mature at different rates, respond to testing at different rates, be influenced by instrumentation in different ways, or show different degrees of regression.

One of psychology's most famous studies is (unfortunately) a good example of a subject selection effect. Known as the "ulcers in executive monkeys" study, it was a pioneering investigation by Joseph Brady in the area of health psychology. Brady investigated the relationship between stress and its physical consequences by placing pairs of rhesus monkeys in adjoining restraint chairs. One monkey, the "executive" (note the allusion to the stereotype of the hard-driving, stressed-out, responsible-for-everything business executive), could avoid mild shocks to its feet that were programmed to occur every 20 seconds by pressing a lever at any time during the interval. For the control monkey (stereotype of the worker with no control over anything), the lever didn't work and it was shocked every time the executive monkey let the 20 seconds go by and was shocked. Thus, both monkeys were shocked equally often, but only one monkey had the ability to control the shocks. The outcome was a stomach ulcer for the executive monkey, but none for the control monkey. Brady then replicated the experiment with a second pair of monkeys and found the same result. He eventually reported data on four pairs of animals (Brady, Porter, Conrad, & Mason, 1958), concluding that the psychological stress of being in command, not just of one's own fate but also of that of a subordinate, could lead to health problems (ulcers in this case).

The Brady study was widely reported in introductory psychology texts, and its publication in *Scientific American* (Brady, 1958) gave it an even broader audience. However, a close examination of Brady's procedure showed that a subject selection confound occurred. Specifically, Brady did not place the monkeys randomly in the two groups. Rather, all eight of them started out as executives in the sense that

they were pretested on how quickly they would learn the avoidance conditioning procedure. Those responding most quickly were placed in the executive condition for the experiment proper. Although Brady didn't know it at the time, animals differ in their characteristic levels of emotionality and the more emotional ones respond most quickly to shock. Thus, he unwittingly placed highly emotional (and therefore ulcer-prone) animals in the executive condition and more laid-back animals in the control condition.

The first to point out the confound was Weiss (1968), whose better-controlled studies with rats produced results the *opposite* of Brady's. Weiss found that those with control over the shock, in fact, developed *fewer* ulcers than those with no control over the shocks.

Attrition

Participants do not always complete the experiment they begin. Some studies may last for a relatively long period of time, and people move away, lose interest, and even die. In some studies, participants may become uncomfortable and exercise their right to be released from further testing. Hence, for any number of reasons, there may be 100 participants at the start of the study and only 60 at the end. This problem sometimes is called subject mortality, or **attrition**. Attrition is a problem because, if particular types of people are more likely to drop out than others, then the group finishing the study is on average made up of different types of people than is the group that started the study. In essence, this is similar to the selection problem because the result is that the group beginning the study is not equivalent to the group completing the study. Note that one way to test for differences between those continuing a study and those leaving is to look at the pretest scores or other attributes at the outset of the study for both groups. If "attriters" and "continuers" are indistinguishable at the start of the study, then overall conclusions at the end of the study are strengthened, even with the loss through attrition.

✓ Self Test 5.3

1. Determined to get into graduate school, Jan takes the GRE nine times. In her first seven attempts, she always scored between 1050 and 1100, averaging 1075. On her eighth try, she gets a 1250. What do you expect her score to be like on her ninth try? Why?
2. How can attrition produce an effect that is similar to a subject selection effect?

This concludes our introduction to the experimental method. The next three chapters will elaborate—Chapter 6 begins by distinguishing between-subjects designs from within-subjects (or repeated measures) designs and describes a number of control problems in experimental research. In particular, it looks at the problems of creating equivalent groups in between-subjects designs, controlling for sequence effects in within-subjects designs, and the biasing effects that result from the fact

that both experimenters and participants are humans. Chapters 7 and 8 look at a variety of research designs, ranging from those with a single independent variable (Chapter 7) to those with multiple independent variables, which are known as factorial designs (Chapter 8).

Chapter Summary

Essential Features of Experimental Research

An experiment in psychology involves establishing independent variables, controlling extraneous variables, and measuring dependent variables. Independent variables refer to the creation of experimental conditions or comparisons that are under the direct control of the researcher. Manipulated independent variables can involve placing participants in different situations, assigning them different tasks, or giving them different instructions. Extraneous variables are factors that are not of interest to the researcher; failure to control them leads to a problem called confounding. When a confound exists, the results could be due to the independent variable or they could be due to the confounding variable. Dependent variables are the behaviors that are measured in the study; they must be defined precisely (operationally).

Manipulated versus Subject Variables

Some research in psychology compares groups of participants who differ from each other in some way before the experiment begins (e.g., gender, age, introversion). When this occurs, the independent variable of interest in the study is said to be selected by the experimenter rather than manipulated directly, and it is called a subject variable. Research in psychology frequently includes both manipulated and subject variables. In a well-controlled study, conclusions about cause and effect can be drawn when manipulated variables are used, but not when subject variables are used.

The Validity of Experimental Research

There are four ways in which psychological research can be considered valid. Valid research uses statistical analysis properly (statistical conclusion validity), defines independent and dependent variables meaningfully (construct validity), and is free of confounding variables (internal validity). External validity refers to whether the study's results generalize beyond the particular experiment just completed.

Threats to Internal Validity

The internal validity of an experiment can be threatened by a number of factors. History, maturation, regression, testing, and instrumentation are confounding factors especially likely to occur in poorly controlled studies that include comparisons

between pretests and posttests. Selection problems can occur when comparisons are made between groups of individuals that are nonequivalent before the study begins (e.g., Brady's ulcers in executive monkeys study). Selection problems also can interact with the other threats to internal validity. In experiments extending over time, attrition can result in a type of selection problem—the small group remaining at the conclusion of the study could be systematically different from the larger group that started the study.

Chapter Review Questions

1. With anxiety as an example, illustrate the difference between independent variables that are (a) manipulated variables and (b) subject variables.
2. Distinguish between Mill's methods of Agreement and Difference, and apply them to a study with an experimental and a control group.
3. Use examples to show the differences between situational, task, and instructional independent variables.
4. What is a confound and why does the presence of one make it difficult to interpret the results of a study?
5. When a study uses subject variables, it is said that causal conclusions cannot be drawn. Why?
6. Describe the circumstances that could reduce the statistical conclusion validity of an experiment.
7. Describe the three types of circumstances in which external validity can be reduced.
8. Explain how the presence of a control group can help reduce the various threats to internal validity. Use history, maturation, or regression as a specific example.
9. Use the Brady study of "ulcers in executive monkeys" to illustrate selection effects.
10. What is attrition and why can it produce interpretation problems similar to subject selection problems?

Applications Exercises

Exercise 5.1—Identifying Variables

For each of the following, identify the independent variable(s), the levels of the independent variable(s), and the dependent variable(s). For independent variables,

identify whether they are manipulated variables or nonmanipulated subject variables. For dependent variables, indicate the scale of measurement being used.

1. In a cognitive mapping study, first-year students are compared with seniors in their ability to point accurately to campus buildings. Some of the buildings are in the center of the campus along well-traveled routes; other buildings are on the periphery of the campus. Participants are asked to indicate (on a scale of 1 to 10) how confident they are about their pointing; the amount of error (in degrees) in their pointing is also recorded.
2. In a study of the effectiveness of a new drug in treating depression, some patients receive the drug while others only think they are receiving it. A third group is not treated. After the program is completed, participants complete the Beck Depression Inventory and are rated on depression (10-point scale) by trained observers.
3. In a Pavlovian conditioning study, hungry animals are conditioned to salivate to the sound of a tone by pairing the tone with food. For some animals, the tone is turned on and then off before the food is presented. For others, the tone remains on until the food is presented. For still others, the food precedes the tone. Experimenters record when salivation first begins and how much saliva accumulates for a fixed time interval.
4. In a study of developmental psycholinguistics, 2-, 3-, and 4-year-old children are shown dolls and asked to act out several scenes to determine if they can use certain grammatical rules. Sometimes each child is asked to act out a scene in the active voice (Ernie hit Bert); at other times, each child acts out a scene in the passive voice (Ernie was hit by Bert). Children are judged by whether or not they act out the scene accurately (two possible scores) and by how quickly they begin acting out the scene.
5. In a study of maze learning, some rats are fed after reaching the end of the maze during the course of 30 trials; others aren't fed at all; still others are not fed for the first 15 trials but are fed for each of the 15 trials thereafter; a final group is fed for the first 15 trials and not fed for the last 15. The researcher makes note of any errors (wrong turns) made and how long it takes the animal to reach the goal.
6. In a helping behavior study, passersby in a mall are approached by a student who is either well dressed or shabbily dressed. The student asks for directions to either the public restroom or the Kmart. Nearby, an experimenter records whether or not people provide any help.

Exercise 5.2—Spot the Confound(s)

For each of the following, identify the independent and dependent variables, the levels of each independent variable, and find at least one extraneous variable that has not been adequately controlled (i.e., that is creating a confound). Use the format illustrated in Table 5.2.

1. A testing company is trying to determine if a new type of driver (club 1) will drive a golf ball greater distances than three competing brands (clubs 2–4). Twenty male golf pros are recruited. Each golfer hits 50 balls with club 1, then 50 more with 2, then 50 with 3, then 50 with 4. To add realism, the experiment takes place over the first four holes of an actual golf course—the first set of 50 balls is hit from the first tee, the second 50 from the second tee, and so on. The first four holes are all 380–400 yards in length, and each is a par 4 hole.
2. A researcher is interested in the ability of schizophrenic patients to judge different time durations. It is hypothesized that loud noise will adversely affect their judgments. Participants are tested two ways. In the "quiet" condition, some participants are tested in a small soundproof room that is used for hearing tests. Those in the "noisy" condition are tested in a nurse's office where a stereo is playing music at a constant (and loud) volume. Because of scheduling problems, locked-ward (i.e., slightly more dangerous) patients are available for testing *only* on Monday and open-ward (i.e., slightly less dangerous) patients are available for testing *only* on Thursday. Furthermore, hearing tests are scheduled for Thursdays, so the soundproof room is available only on Monday.
3. An experimenter is interested in whether memory can be improved if people use visual imagery. Participants (all females) are placed in one of two groups—some are trained in imagery techniques, and others are trained to use rote repetition. The imagery group is given a list of 20 concrete nouns (for which it is easier to form images than abstract nouns) to study, and the other group is given 20 abstract words (ones that are especially easy to pronounce, so repetition will be easy), matched with the concrete words for frequency of general usage. To match the method of presentation with the method of study, participants in the imagery group are shown the words visually (on a computer screen). To control for any "compu-phobia," rote participants also sit at the computer terminal, but the computer is programmed to read the lists to them. After hearing their respective word lists, participants have 60 seconds to recall as many words as they can in any order that occurs to them.
4. A social psychologist is interested in helping behavior and happens to know two male graduate students who would be happy to assist. The first (Ned) is generally well dressed, but the second (Ted) doesn't care much about appearances. An experiment is designed in which passersby in a mall will be approached by a student who is either well-dressed Ned or shabbily dressed Ted. All of the testing sessions occur between 8 and 9 o'clock in the evening, with Ned working on Monday and Ted working on Friday. The student will approach a shopper and ask for a dollar for a cup of coffee. Nearby, the experimenter will record whether or not people give money.

Exercise 5.3—Operational Definitions (Again)

In Chapter 3, you first learned about operational definitions and completed an exercise on the operational definitions of some familiar constructs used in psychological research. In this exercise, you are to play the role of an experimenter designing a study. For each of the four hypotheses:

a. identify the independent variable(s), decide how many levels of the independent variable(s) you would like to use, and identify the levels;
b. identify the dependent variable in each study; and
c. create operational definitions for your independent and dependent variables.

1. People will be more likely to offer help to someone in need if the situation unambiguously calls for help.
2. Ability to concentrate on a task deteriorates when people feel crowded.
3. Good bowlers improve their performance in the presence of an audience, whereas average bowlers do worse.
4. Animals learn a difficult maze best when they are moderately aroused. They do poorly in difficult mazes when their arousal is low or high. When the maze is easy, performance improves steadily from low to moderate to high arousal.

Answers to the Self Tests:

✓ **5.1.**

1. IVs = problem difficulty and reward size
   DV = number of anagrams solved
2. Extraneous variables are all of the factors that need to be controlled or kept constant from one group to another in an experiment; failure to control these variables results in a confound.
3. Frustration could be manipulated as an IV by having two groups, one allowed to complete a maze, and the other prevented from doing so. It could also be an extraneous variable being controlled in a study in which frustration was avoided completely. It could also be what is measured in a study that looked to see if self-reported frustration levels differed for those given impossible problems to solve, whereas others are given solvable problems.

✓ **5.2.**

1. As a manipulated variable, some people in a study could be made anxious ("you will be shocked if you make errors"), and others not; as a subject variable, people who are generally anxious would be in one group, and low anxious people would be in a second group.
2. Manipulated → the viewing experience shown to children.
   Subject → gender.
3. Internal → the study is free from methodological flaws, especially confounds.
   External → results generalize beyond the confines of the study.
4. Ecological.

✓ **5.3.**

1. Somewhere around 1275; regression to the mean
2. If those who drop out are systematically different from those who stay, then the group of subjects who started the study will be quite different from those who finished.