Introduction to Experimental Research

5

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 factors of interest (the independent variables), controlling all other factors (extraneous variables), and measuring outcomes (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:

- Describe the impact of Robert Woodworth's 1938 *Experimental Psychology* on the way psychologists define an experiment.
- Define a manipulated independent variable, and identify examples of 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 cofounds create serious problems for interpreting the results of an experiment.
- Identify independent and dependent variables, given a brief description of any experiment.
- Distinguish between independent variables that are manipulated variables and those that are 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.
- Distinguish between the internal and external validity of a study.
- Describe the various factors affecting an experiment's external validity.
- 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 published his *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 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 are now reading, is 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 in psychology today. In particular, he contrasted experimental with correlational research, a distinction now well known by research psychologists.

The defining feature of the experimental method was the manipulation of what Woodworth (1938) called an "independent variable," which affected 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" (p. 2). Although Woodworth did not invent these terms, he was the first to use them as they are used routinely today.

While the experimental method manipulates independent variables, the correlational method, according to Woodworth (1938), "[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" (p. 3). You will learn more about correlational research in later chapters. 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. After all, as psychologists we seek to discover the causes of behavior; thus, the experimental method is the best tool to help up understand those causes and what effects they have on behavior.

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 factor (or factors), holds all other factors constant, and observes the results of the variation. The factors under the control of the

experimenter are called *independent variables*, the factors being held constant are the *extraneous variables*, and the behaviors measured are called *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 19th 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 found in physics, it could do just as well as some 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 later discussion of correlation. The methods were those of "Agreement" and "Difference" (relevant for this chapter), and of "Concomitant Variation" (relevant for correlation, covered in Chapter 9).

Taken together, the methods of Agreement and Difference enable us to conclude, with a high degree of confidence (but not absolute certainty), that some factor, *X*, causes some result, *Y*. 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 *X* does not occur and Y also does not occur, then *X* is *necessary* for *Y* to occur—"if no *X*, then no *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 *Y* to happen.

To make this more concrete, suppose we are trying to determine if watching violent TV causes children to become 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 (*X*), then aggression (*Y*) 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. This satisfies the Method of Difference. If the child does not watch violent TV, aggression does not occur ("If no *X*, then no *Y*"). This combined outcome (Joint Method) would establish that watching TV causes aggression in children.

It is important to note that in the real world of research, the conditions described in these methods are never fully met. It is impossible to identify and measure what happens every time a child watches TV. Rather, the best one can do is to observe systematically as many examples as possible, under controlled conditions, and then draw conclusions with a certain amount of confidence, based on some form of statistical analysis. 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, whereas 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 what Woodworth called the **independent variable**: It is the factor of interest to the experimenter, the one being studied to see if it will influence behavior (the "watching violent TV" in the John Stuart Mill example). It is sometimes called a *manipulated* variable or factor because the experimenter has complete control over it and is creating the situations research participants will encounter in the study. As you will see, the concept of an independent variable can also be stretched to cover *non-manipulated* or *subject variables*, but, for now, let us consider only those independent variables that are under the experimenter's total control. We will refer to them as *manipulated independent variables*.

All manipulated 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 caffeine on reaction time. Such a study requires at least two dosage levels of caffeine in order to make a comparison. This study would be described as an experiment with "amount of caffeine consumed" as the manipulated independent variable and two different dosages as the two levels of the independent variable. You could also say the study has two conditions: the two dosages. 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 learn in Chapter 7.

As you recall from Chapter 3, experimental research can be either basic or applied, and it can be conducted either in the laboratory or in the field. 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 the laboratory, including both experimental studies and studies using non-experimental methods.

Varieties of Manipulated Independent Variables

The range of factors that can be used as manipulated 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, task, and instructional variables.

Situational variables are 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 subjects 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 independent variable is the number of potential helpers on the scene besides the participant, and the levels are zero, three, and six bystanders. Thus the experimenter has created three situations.

Sometimes, experimenters vary the type of task performed by subjects. One way to manipulate **task variables** is to give participants different kinds of problems to solve. For instance, research on the psychology of reasoning often involves giving people different kinds of logic problems to determine the kinds of errors people tend to make. Similarly, mazes can differ in degree of complexity, different types of illusions could be presented in a perception study, and so on.

Instructional variables are manipulated by telling different groups to perform a particular task in different ways. For example, students 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.

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—a task variable. Finally, an instructional variable could manipulate motivation by telling participants they will earn either $1 or $5 for completing the puzzles.

Control Groups

In some experiments, the independent variable is whether or not some experimental condition occurs. Some subjects get the treatment condition, and others do not. 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 (they will probably be shown a non-violent TV program). 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 the control group participants do not get the experimental treatment. As you recall from Box 5.1, the conditions of the experimental group (shown violent TV) satisfy Mill's Method of Agreement (if violent TV, then aggression) and the control group (not shown violent TV) 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 Mill's 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 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 only in research when it is important to have a comparison with some baseline level of performance. For example, suppose you were interested in the construct "sense of direction" and wanted to know whether a 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 that did not get any training. On the other hand, if your empirical question concerns gender differences in sense of direction, the comparison will be between a group of male subjects and a group of female subjects; 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. For an example of a study comparing simple experimental and a control groups, consider this interesting study about superstition.

Research Example 6—Experimental and Control Groups

Is there such a thing as good luck? As a golfer, will you putt better if you think you're using a lucky ball? In a manual dexterity game, will you do better if others say they have their fingers crossed for you? Will you do better on a memory or an anagram task if you have your lucky charm with you? The answer to all these questions appears to be *yes*, according to a simple yet clever set of studies by Damisch, Stoberock, and Mussweiler (2010). In each of four studies, subjects in an experimental group were given reason to think they might be lucky; those in a control group were not given any reason to think that luck would occur. In the first study, subjects were asked to make 10 putts of 100 cm (just over 3 feet). Experimental group subjects were handed a golf ball and told, "Here is your ball. So far, it has turned out to be a lucky ball" (p. 1015); control group subjects were told, "This is the ball everyone has used so far" (p. 1015).

Here are the means (*M*) (average number of putts made) and standard deviations (*SD*) for the two groups:

$$\text{Experimental group}(\text{lucky ball}): \quad M = 6.42 \quad SD = 1.88$$
$$\text{Control group}(\text{no mention of luck}): \quad M = 4.75 \quad SD = 2.15$$

Recall from Chapter 4 that to achieve a significant difference between two groups in null hypothesis significance testing, there should be a noticeable difference between the mean scores for the groups and a relatively small amount of variability within each group. That happened in this case—the experimental group holed significantly more putts than the control group and the standard deviations are small; in addition, the effect size was determined to be large (Cohen's *d* equaled an impressive .83).

Damisch et al. (2010) reacted to the putting outcome as we suspect you might have—they weren't quite sure they believed such a minor thing ("you have a lucky ball") could make such a big difference. So they did three more experiments, all using the same basic design—experimental group subjects being given some indication they would have luck on their side, control group subjects given no such indication. Manipulating "luck" as the independent variable worked in all the experiments. Subjects believing luck would be with them outperformed controls on a manual dexterity task, a memory task, and a problem-solving task. In the dexterity task, the experimenter simply told subjects in the experimental group "I'll have my fingers crossed for you." In the memory and problem-solving tasks, experimental subjects were allowed to keep a "lucky charm" with them as they did the experiment. All three of these subsequent studies replicated the results of the one that measured putting performance. Of course, Damisch and her colleagues did not conclude that luck existed as an explanation for the results. Rather, based on some other measures they took, they concluded that those in the experimental group simply believed they would do well and, as a result of their enhanced "self-efficacy," did the kinds of things that might be expected to improve performance (e.g., concentrate harder than those in the control groups).

The concept of a control group has a long history. As just one example, Francis Galton had it in mind for his famous study on the efficacy of prayer that you learned about in Chapter 1. As you recall, he concluded that prayers had no measurable effect. When planning the study, Galton (1872) argued that for the study to be of any value, "prudent pious people must be compared with prudent materialistic people and not with the imprudent nor the vicious" (p. 126). That is, those who prayed (i.e., the pious, an experimental group) were to be compared with those who did not pray (i.e., the materialistic, a control group), but it was important that the two groups be alike in all other ways (e.g., prudent). You will learn in Chapter 6 that Galton was advocating the idea of creating comparable groups through a procedure of *matching*—selecting two groups carefully so that they are alike in all important ways except for the conditions being compared.

Controlling Extraneous Variables

After manipulating independent variables, the second feature of the experimental method is that the researcher tries to control **extraneous variables**. These are variables that are not of interest to the researcher but that might influence the behavior being studied if not controlled properly. As long as these are held constant, they present no danger to the study. In the "putt your lucky ball" study, for instance, putts attempted by subjects in both groups were always of the same length. In Galton's prayer study, it was important for both the pious and the non-pious to be equally "prudent." If a researcher fails to control extraneous variables, they can systematically influence the behavior being measured. The result is called *confounding*. A **confound** is any uncontrolled extraneous variable that co-varies with the independent variable and could provide an alternative explanation of the results. That is, a confounding variable changes in the same way

that an independent variable changes (i.e., they co-vary) and, consequently, its effect cannot be distinguished 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 know which variable explains the results. Thus, results from studies with confounds are uninterpretable.

To illustrate confounding, consider a verbal learning experiment in which a researcher wants to show that students who try to learn a large amount of course material all at once don't do as well as those who spread their studying over several sessions—that is, massed practice (e.g., cramming the night before an exam) is predicted to be inferior to distributed practice. Three groups of students are selected, and each group is given the same chapter 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. Table 5.1 shows the design. The results show that subjects in Group 3 score the highest, followed by those in Group 2; Group 1 subjects do not do well at all. On the basis of this outcome, the researcher concludes that distributed practice is superior to massed practice. Do you agree with this conclusion?

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 3s in each column equal the number of hours spent studying a chapter of a psychology text.

You probably don't (we hope) because there are two serious confounds in this study, and both should be easy for you 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 during the week (3, 6, or 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). In other words, the two factors co-vary perfectly. The way to describe this situation is to say "the distribution of practice is confounded with total study hours." A 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, the distribution of practice is confounded both with total study hours and with retention interval. Each confound by itself could account for the results.

Table 5.2 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 methodological controls. If they are not kept constant, then confounding may occur. 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 combination. To correct the confound problem in this case, you must ensure the middle two columns each have the same terms in them. Thus, instead of 3, 6, or 9 hours for the first extraneous variable (EV#1), the total number of hours spent studying should be the same. Likewise, instead of 1, 2, or 3 days for EV#2, the number of days in the retention interval should be the same. Table 5.3 shows you one way to control for what is confounded in Table 5.1. As you can see, total study time and retention interval are held constant.

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 = independent variable.
EV = extraneous variable.
DV = dependent variable.

Table 5.3 Eliminating Confounds in a Hypothetical Distribution of Practice Experiment

	Monday	Tuesday	Wednesday	Thursday	Friday
Group 1	—	—	—	3	Exam
Group 1	—	—	1.5	1.5	Exam
Group 1	—	1	1	1	Exam

A problem students sometimes have with understanding confounds is that they tend to use the term whenever they spot something in a study that might appear to be a flaw. For example, suppose the distribution of practice study included the statement that all the subjects in the study were between 60 and 65 years old. Some students reading the description might think there's a confound here that concerns age. What they really mean is they believe a wider range of ages ought to be used. They could be right, but age in this case is not a confound. Age would be a confound *only* if subjects in the three groups were of three *different* age ranges. If those in Group 1 were aged 60–65, those in Group 2 were 30–35, and those in Group 3 were 18–22, then a confound would exist. Group differences in the results could be due to the independent variable or to age; those in Group 3 might do better because their studying has be spread out, but they might do better simply because they are younger.

Learning to be aware of potential confounding factors and building appropriate ways to control for them is a scientific thinking skill that is difficult to develop. Not all confounds are as obvious as the massed/distributed practice example. We'll encounter the problem occasionally in the remaining chapters and address it again shortly when we discuss 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 the behavior that is the measured outcome of a study. 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 example from Box 5.1 on Mill's Joint Method), 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. In Chapter 3's discussion of operational definitions, we have already seen that empirical questions

cannot be answered unless the terms are defined with some precision. When an experiment is designed, therefore, one key component concerns the operational definitions for the behaviors to be measured as dependent variables. Unless the behaviors are defined precisely in terms of their measurement, direct 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 valid and reliable. 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 groups in the study are so high that no difference can be determined between conditions. For example, this can happen 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, usually because the task is too difficult for everyone, once again producing a failure to find any differences between conditions.

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 experimenter 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 (instructional independent variable) that they will experience shocks, either mild or painful, when they make errors on a simulated driving task. Anxiety could also be a factor that must 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 video the students in one group without doing so in the other group as well. If everyone is videoed, then the level of anxiety created by that factor (video recording) 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 versus 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 characteristically having more of it than others. This last possibility leads to our 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.

Subject Variables

Up to this point, the term *independent variable* has meant a factor directly manipulated 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 made between groups of people who differ from each other in ways other than those directly manipulated by the researcher. Factors that are not directly manipulated by an experimenter are referred to variously as ex post

facto variables, natural group variables, participant variables, or **subject variables**, which will be our focus here. Subject variables are 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 conditions of the experiment by virtue of the characteristics they already have.

To see 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 people who tend to be anxious all the time (as determined ahead of time, perhaps, by a personality test for anxiety proneness). The second group would include more relaxed 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* upon entering the study.

Some researchers, true to Woodworth's original use of the term, prefer to reserve the term *independent variable* for variables directly manipulated by the experimenter. Others are willing to include subject variables as examples of a particular type of independent variable because the experimenter has some degree of control over them by virtue of the decisions involved in selecting them in the first place and because the statistical analyses will be the same in both cases. We take this latter position and will use the term *independent variable* in the broader sense. However, whether the term is used broadly (manipulated or subject) or narrowly (manipulated only), it is important that you understand the difference between a manipulated variable and a non-manipulated, subject variable both in terms of how the groups are formed in the study, and the kinds of conclusions that can be drawn from them.

Research Example 7—Using Subject Variables

One common type of research using subject variables examines differences between cultures. Ji, Peng, and Nisbett (2000) provide a nice example. They examined the implications of the differences between people raised in Asian cultures and those raised in Western cultures. In general, they pointed out that Asian-Americans, especially those with families 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 subject variables: culture and gender. For their dependent measure, they chose performance on 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 participants from Asian cultures would be more likely to be field dependent than those from Western cultures. They also hypothesized greater field dependence for women, a prediction based on a typical finding in RFT studies. The standard RFT procedure was used in the same way it was used in past studies. So, in the replication terms you learned about in Chapter 3, part of this study (gender) involved direct replication and part (culture) involved conceptual replication.

Because the undergraduate population of the University of Michigan (where the research was conducted) included a large number of people originally from East Asia, Ji et al. (2000) were able to complete their study using students enrolled in general psychology classes there (in a few pages you'll be learning about university "subject pools"). They compared 56 European-Americans with 42 East Asian-Americans (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 each group had about an equal number of male and female participants.

As you can see from Figure 5.1, the results supported both hypotheses (larger error scores on the *Y*-axis indicated a greater degree of field dependence). The finding about women being more field dependent than men was replicated and that difference occurred in both cultures. In addition, the main finding was the 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 dependence of those from Asian cultures 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. Each subject did the RFT task 16 times

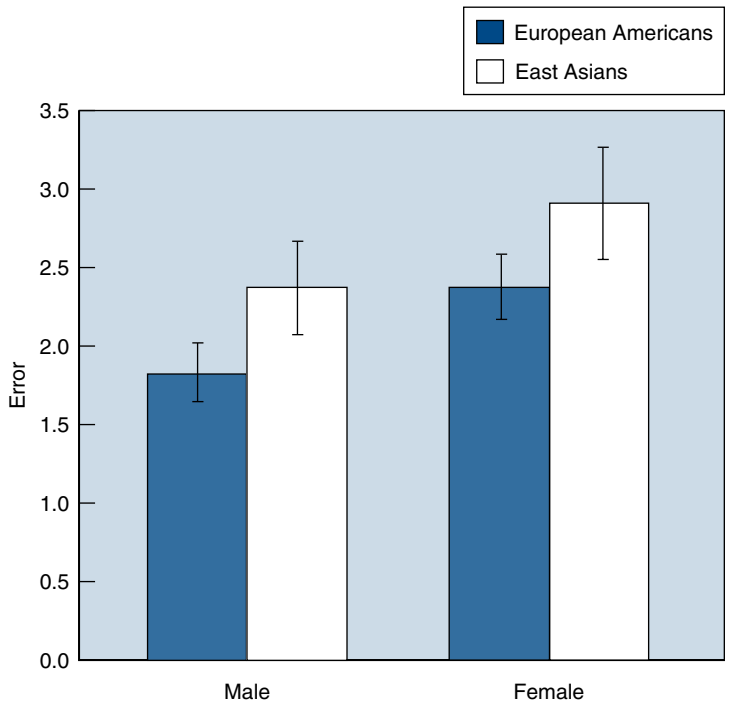

FIGURE 5.1
Gender and cultural differences in the rod and frame test, from Ji, Peng, and Nisbett's (2000) cross-cultural study.

and, on average, 1.2 of the 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.

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 somehow false). Studies using independent variables that are *subject* variables are occasionally called *ex post facto studies*, *natural groups studies*, or *quasi experiments* (*quasi* meaning "to some degree" here).[1] Studies often include both manipulated and subject independent variables, as you will learn in Chapter 8. Being aware of subject variables is important because they affect the kinds of conclusions that can be drawn.

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 some degree of confidence; with subject variables, causal conclusions cannot be drawn. The reason has to do with the amount of control held by the experimenter.

With manipulated variables, the experiment can meet the criteria listed in Chapter 1 for demonstrating causality. The independent variable precedes the dependent variable and can be considered the most reasonable explanation for the results, assuming no confounds are present. In other words, if you vary one factor and successfully hold all else constant, the results can be attributed *only* to the factor varied.

When using subject variables, however, the experimenter can vary a factor (i.e., select participants having certain characteristics) but cannot hold all else constant. Selecting participants who are high or low on anxiety proneness does not guarantee the two groups will be equivalent in other ways. In fact, they might differ in several ways (e.g., self-confidence, tendency to be depressed) that could influence the outcome of the study. When a difference between the groups occurs in this type of study, we cannot say the differences were *caused* solely by the subject variable. In terms of the conditions for causality, although we can say the independent variable precedes the dependent variable, 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.

An example from social psychology might help 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 subjects an achievement test. By providing different kinds of false feedback about their performance on the test, either positive or negative, self-esteem could be raised or lowered temporarily. The participants could then be asked to do 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 valid and reliable self-esteem test and select those who score high or low on the measure as the participants for the two groups. Self-esteem in this case is a subject variable; half of the participants will

[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 will be elaborated in Chapter 11.

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

naturally have low self-esteem, while the other half will naturally have high self-esteem. As in the first study, these two groups of people then 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 high self-esteem is the direct cause of the helping behavior; all you can say is that people with high self-esteem are more likely to help than those with low self-esteem. Your conclusion would then be limited to making educated guesses about the reasons why this might be true because these participants may differ from each other in other ways unknown to you. For instance, people with high self-esteem might have had prior experience in volunteering, and this experience might have had the joint effect of raising or strengthening their characteristic self-esteem and increasing the chances they would 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 Chapters 9 and 10, this difficulty in interpreting research with subject variables is exactly the same problem encountered when trying to draw conclusions from correlational research (Chapter 9) and quasi-experimental research (Chapter 10).

Returning for a moment to the Ji et al. (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." In their words, "European Americans made fewer mistakes on the RFT than East Asians, . . . [and] men made fewer mistakes than women" (p. 950).

Before moving on to the discussion of the validity of experimental research, read Box 5.2. It identifies the variables in a classic study you probably recall from your general psychology course—one of the so-called Bobo doll 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.

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 (see Figure 5.2). The response will be immediate recognition and a brief description along the lines of "Oh, yes, the studies showing that children will punch an inflated doll if they see an adult doing it." A description of one of these studies is a good way to clarify 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 one control group.

Experimental Group 1: real-life aggression (children directly observed an adult model aggressing against a 5-foot-tall 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 non-manipulated independent variable (subject variable) was gender. Male and female children from the Stanford University Nursery School (mean age = 52 months) were the participants in the study. (Actually, there was also

(continued)

BOX 5.2 (CONTINUED)

The Drs. Nicholas and Dorothy Cummings Center for the History of Psychology, The University of Akron.

FIGURE 5.2
One of Bandura's Bobo dolls, donated by Bandura to the Center for the History of Psychology at the University of Akron.

another manipulated variable; participants in Groups 1 and 2 were exposed to either a same-sex or opposite-sex 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 a 3-foot-tall Bobo doll,* thereby giving the children the opportunity to be aggressive themselves.

Controlling Extraneous Variables

Several possible confounds were avoided. First, whenever a child was put into the room with the toys to see if aggressive behavior would occur, the toys were always arranged in exactly the same way "in order to eliminate any variation in behavior due to mere placement of the toys in the room" (Bandura et al., 1963, p. 5). 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, *all* of the children had an approximately equivalent increase in their degree of emotional arousal just prior to the time they were given the opportunity to act aggressively. In other words, any differences in aggressiveness could be attributed to the imitative effects and not to any emotional differences between the groups.

*Notice that the adults hit a *5-foot tall* Bobo, but the children were given the opportunity to hit a smaller doll, a *3-foot* Bobo. This is a nice design feature in the study. Can you see why?

Measuring Dependent Variables

Several measures of aggression were used in this study. Aggressive responses were categorized as imitative, partially imitative, or non-imitative, 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. Non-imitative aggression included shooting darts from an available dart gun at targets other than Bobo and acting aggressively toward other objects in the room.

Briefly, the results of the study were that children in Groups 1, 2, and 3 showed significantly more aggressive behavior 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 5.3 summarizes the results.

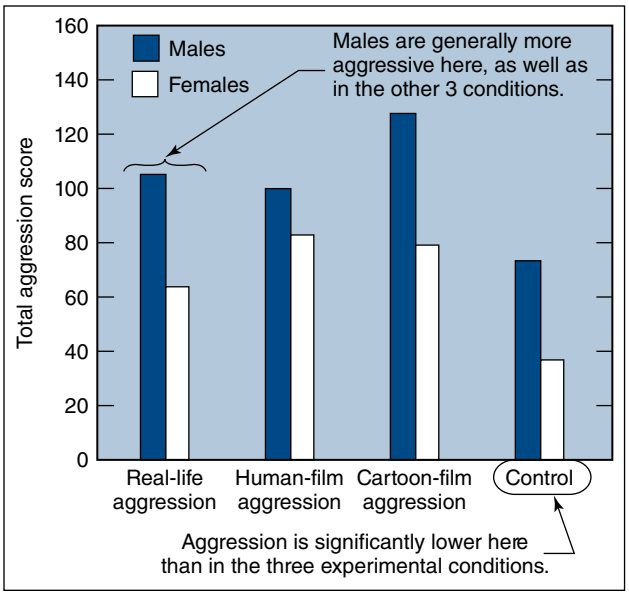

FIGURE 5.3
Data from Bandura, Ross, and Ross's Bobo study (1963) of the effects of imitation on aggression.

The Validity of Experimental Research

Chapter 4 introduced the concept of validity in the context of measurement. The term also applies to research methodology 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 it is supposed to provide. This section of the chapter introduces four types of validity, following the scheme first 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, of major importance, 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 conclusion 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, but the researcher mistakenly uses an

analysis appropriate only for interval or ratio data. Another factor reducing 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 exists (i.e., H_0 should be rejected) but low reliability results in a failure to find that effect, the outcome is a Type II error, which reduces the statistical conclusion validity.

Careful researchers decide on the statistical analysis at the same time they plan the experimental design. In fact, no experiment should ever be designed without thought given to how the data will be analyzed.

Construct Validity

Chapter 4 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, referring 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 that could be chosen. If the study used violent cartoon characters (e.g., Tom and Jerry) 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 pounding each other into submission. 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 led to believe they are delivering electric shocks to another person) than others (e.g., horn honking by frustrated drivers). As was true in our earlier discussion of construct validity, 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 and theoretically meaningful pattern.

External Validity

Experimental psychologists have been criticized for knowing a great deal about undergraduate students 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 undergraduates fits here. As we saw in Chapter 2, part of the debate about the appropriateness of animal research has to do with how well this research provides explanations relevant for human behavior. Concerning undergraduates, recall that Milgram deliberately avoided using college students, selecting adults from the general population as subjects for his obedience studies. The same cannot be said of most psychologists, however. In an analysis of all empirical papers in six top-tier psychology journals published between 2003 and 2007,

Arnett (2008) reported that 68% of participants were from the United States alone, and, including the U.S., 96% of participants were from Western industrialized countries. Furthermore, 67% of U.S. study participants and 80% of participants in other countries were undergraduate students. Henrich, Heine, and Norenzayan (2010) suggested that undergraduate students are "a truly unusual group" (p. 61) and "outliers within an outlier population" (p. 78) referring to them as WEIRD – or "people from Western, Educated, Industrialized, Rich, and Democratic societies" (p. 61). They argued that researchers must be extremely careful when generalizing their results to human beings in general when their sample (i.e., likely college undergraduates) represents a very small subset of the global population. However, Sears (1986) pointed out that many research areas (e.g., perception, memory, and attention) 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. One prominent memory researcher (Roediger, 2004) went so far as to argue that college students were the *ideal* subjects for his research: "Millions of years of evolution have designed a creature that is a learning and memorizing marvel. Students in my experiments have also been carefully selected through 12 or more years of education before they get to my lab. The world could not have arranged a more ideal subject" (p. 46). Some special ethical considerations apply when using college students, especially when recruiting them from introductory psychology courses. Box 5.3 lists some guidelines for using a "subject pool" ethically.

BOX 5.3 ETHICS—Recruiting Participants: Everyone's in the Pool

Most research psychologists are employed by colleges and universities and consequently are 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 a **subject pool** or 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 probably had this experience when you took your introductory psychology course. At a large university, if 800 students take the intro course each semester and each student signs up for three studies, 2,400 participants are 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, 2007). Ideally, students acquire insight 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. A study by Sieber and Saks (1989), for example, found evidence that 89% of 366 departments surveyed had pools that failed to meet at least one of the APA's recommendations (below). Critics sometimes 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. On the other hand, a study by Trafimow, Madson, and Gwizdowski (2006) found that, when given a choice between research participation and a brief paper that was described as requiring the same amount of effort as participation, most students opted for participation, and a substantial number (43.5% of those surveyed) indicated that participation in research while in introductory psychology had increased their interest in psychology.

Although there is 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 the memory chapter 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 community volunteer work, or (c) attend

(continued)

BOX 5.3 (CONTINUED)

a research presentation by a visiting scholar and write a brief summary of it (Kimmel, 2007; McCord, 1991). Some studies have shown that students generally find research participation valuable, especially if researchers make an attempt to tie the participation to the content of the introductory psychology course (e.g., Landrum & Chastain, 1999; Leak, 1981).

The APA (1982, pp. 47–48) has provided guidelines for 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 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 subject pool policies.

Testing only undergraduate students is only one example of the concern about generalizing to other groups. Another has to do with gender. Some of psychology's most famous research has been limited by studying 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 became 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 overlooked important gender differences in thinking patterns and in how moral decisions are made. Males may place the highest value on individual rights, but females tend to value the preservation of individual relationships. Hence, girls responding to some of Kohlberg's moral dilemmas might not seem to be as morally "advanced" as boys, but this is due to the bias of the entire model because Kohlberg sampled only boys, 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). Hence, research conclusions based on just one culture might not be universally applicable. Again, Henrich et al. (2010) reviewed many differences between WEIRD (Western, Educated, Industrialized, Rich, and Democratic) people compared to other populations, and many psychological theories are less clear when one considers other cultures or populations. For example, Research Example 7 found a cultural difference in field dependence. As another 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 all research in psychology should make cross-cultural comparisons? No. It just means conclusions must be drawn cautiously and with reference to the group studied in the research project.

Other Environments

Besides generalizing to other types of individuals, externally valid results are applicable to other 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 (Chapter 3) 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 20th century, memory research occurred largely in the laboratory, where countless undergraduate students 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 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). And as you might guess, the concept of ecological validity found its way into many areas of psychology, not just cognitive psychology. In social psychology, the topic of interpersonal attraction, for instance, Finkel and Eastwick (2008) hit on the creative idea of using a speed-dating format for their research. As you probably know, in speed dating, couples pair off for a brief period of time and introduce themselves to each other and then move on to subsequent pairings. If the event includes 10 men and 10 women, for instance, the men and women will be paired together randomly, interact for perhaps 5 minutes, and then a new pairing will occur, leading to another 5-minute event, and so on, until each man has interacted with each woman. Then (or sometimes after each pairing, to avoid memory problems) all 20 people complete a survey indicating how they perceived each of the other persons they met, and organizers arrange for pairs that seem attracted to each other to meet again on their own. Finkel and Eastwick argued that the procedure is ideal for research—there is a high degree of control over the interactions (e.g., each lasts a fixed amount of time), there is a great opportunity to collect mountains of data (e.g., eye contact data can be matched to attractiveness judgments), and the procedure has a high degree of ecological validity. It is a real-life event populated with people who are genuinely interested in meeting others and

perhaps developing a continuing relationship, and it resembles other kinds of circumstances where couples meet for the first time (e.g., a blind date).

Other Times

The third way in which external validity is sometimes questioned has to do with the longevity of results or the historical era during which a particular experiment was completed. An example is the study used to illustrate an ordinal scale in Chapter 4. Korn, Davis, and Davis (1991) found that department chairpersons ranked B. F. Skinner first on a list of top 10 contemporary psychologists. But Skinner had died just the year before and his lifetime achievements were highly visible at the time. Replicating that study 25 years later might very well produce a different outcome. As an older example, 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 at a time 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 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 historical context.

A Note of Caution about External Validity

Although external validity has value under many circumstances, it is important to point out that it is not often a major concern in the design of a research project. 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 the so-called artificial conditions, which 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 the word was on the list. 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, even if they are wrong), 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 much research that continues to explore the theoretical basis for false memory, determining, for instance, the limits of the phenomenon and exactly how it occurs. Eventually, that research will produce a body of knowledge that comprehensively explains the false memory phenomenon.

In summary, the external validity of a research finding increases as it applies to other people, places, and times. But must researchers design a study that includes many groups of people, takes place in several settings, including "realistic" ones, and is repeated every decade? Of course not. External validity is not determined by an individual research project; it accumulates over time as research is replicated in various contexts. 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 are 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. 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 interpersonal attraction during speed dating was used to illustrate which form of validity?

Threats to Internal Validity

Any uncontrolled extraneous factor (i.e., the confounds you learned about earlier in the chapter) can reduce a study's internal validity, but a number of problems require special notice (Cook & Campbell, 1979). These problems or threats to internal validity are notably dangerous when control groups are absent, an issue that sometimes occurs in an applied form of research called *program evaluation* (see Chapter 11). Many of these threats occur in studies that extend over a period during which several measures are taken. For example, participants might receive a pretest, an experimental treatment of some kind, and then a posttest, and maybe even a follow-up test. Ideally, the treatment should produce a positive effect that can be assessed by observing changes from the pretest to the posttest, changes that are maintained in the follow-up. A second general type of threat occurs when comparisons are made between groups said to be "nonequivalent." These so-called subject selection problems can interact with the other threats to internal validity.

Studies Extending Over Time

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 over time as the result of some experience (a course, a program, and memory training). To judge whether change occurred, one procedure is to evaluate people prior to the experience with a **pretest**. Then, after the experience, a **posttest** measure is taken. Please note that although we will be using pretests and posttests to illustrate several threats to internal validity, these threats can occur in any study extending over time when participants are tested multiple times, whether or not pretests are used.

The ideal outcome for the examples we've just described is that, at the end of the period for the study, 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. A typical research design includes pretests and posttests and compares experimental and control groups:

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

If this type of procedure occurs without a control group, there are several threats to internal validity. For example, suppose we are trying to evaluate the effectiveness of a college's program to help incoming students who suffer from test anxiety—that is, they have decent study skills and seem to know the material, but they are so anxious during exams they don't perform well on them. During freshman orientation, first-year students complete 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. Some incoming students who score very high (i.e., they have the greatest need for the program) are asked to participate in the college's test anxiety program, which includes relaxation training, study skills training, and other techniques. Three months later, these students 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? We hope you can see several ways of interpreting this outcome and that it is not at all certain the program worked.

History and Maturation

Sometimes, an event outside of the study occurs between pre- and post-testing that produces large changes unrelated to the treatment program itself; 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 will henceforth be graded on a pass/fail basis. Furthermore, suppose that 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 students at the very start of their college careers, so changes in scores 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, or developmental changes that occur with the passage of time, is always a concern whenever a study extends over time.

Notice that if a control group is used, the experimenter can account for the effects of both history and maturation. These potential threats could 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, 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 the Mean

To regress is to go back, in this case in the direction of a mean (average) score. Hence, the phenomenon described here is sometimes called **regression to the mean**. In essence, it refers to the fact that if the first score from a subject is an extreme score, then the second or third score from the same person will be closer to whatever the mean is for the larger set of scores. This is because, for 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 5.4. Most of the scores center around the mean, so, if you make a random selection, you'll most likely choose a score near the mean (*X* on the left-hand graph of Figure 5.4). However, suppose you 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 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?

Our guess is you've chosen alternative "c," which means you understand the basic concept of regression to the mean. To take a more concrete example (refer to the right-hand graph of Figure 5.4), 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, you've probably chosen *c*, further convincing yourself that you get the idea of the regression phenomenon. But what does this have to do with our study about test anxiety?

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 selected because on the pretest they scored very high for anxiety. On the posttest, their anxiety scores might be lower than on the pretest, but the change in scores could be due to regression to the mean, at least in part, rather than the result of the anxiety improvement program. Once again, a control group of equivalent high-anxiety participants would enable the researcher to control for regression to the mean. For instance, the following outcome would suggest regression to the

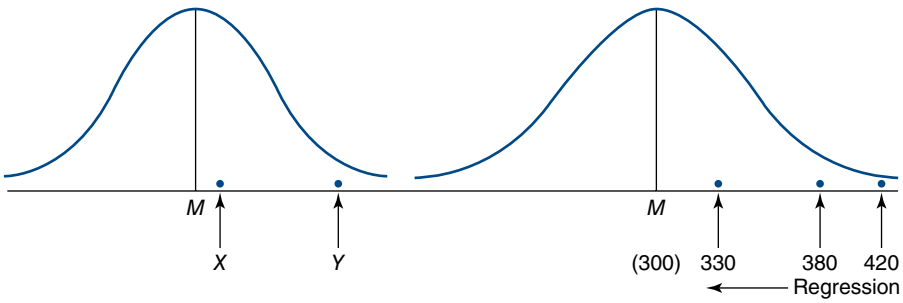

FIGURE 5.4
Regression to the mean.

mean might be involved,[3] but the program nonetheless seemed to have an effect. Can you see why this is so?

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

Regression to the mean 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 11 as an example of problems involved in assessing large-scale, federally supported programs.

Testing and Instrumentation

Testing is considered a threat to internal validity when the mere fact of taking a pretest has an effect on posttest scores. There could be a practice effect of repeated testing, or aspects of the pretest could sensitize participants to something about the program. For example, if the treatment program is a self-paced, computerized psychology course, the pretest would be a 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 the measurement instrument changes from pretest to posttest. In the self-paced 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 to the mean, 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 be less likely to have that opportunity.

Participant Problems

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

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 in all ways except for the independent variable. In the next chapter, you will learn how these equivalent groups are formed through random assignment or matching. If the groups are not equivalent, then **subject selection effects**

[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 be ruled out.

might occur. For example, suppose two sections of a 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, and perhaps better students) 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 and his colleagues investigated the relationship between stress and its physical consequences by placing pairs of rhesus monkeys in adjoining restraint chairs (Brady, Porter, Conrad, & Mason, 1958). 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 et al. then replicated the experiment with a second pair of monkeys and found the same result. They eventually reported data on four pairs of animals, concluding 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 research 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 a subject selection confound. 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 learning 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 selection 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, 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, which affects the external validity of the study. It is a particular problem when one condition in the study has a higher attrition rate than another condition. Now you have a confound and a clear threat to internal validity. In either case, in the final analysis, 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 it 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.

A Final Note on Internal Validity, Confounding, and External Validity

As you recall from Chapter 1, one of the goals of this text is to make you critical consumers of psychological research—that is, we hope you will be able to read research and spot any flaws. Sometimes, students have a tendency to overuse the term *confound*, reporting that a study is "confounded" when they spot a flaw that is, in fact, not a confound. Specifically, they tend to confuse the issues of internal and external validity. Remember, *internal validity* refers to the methodological soundness of a study—it is free from confounds. *External validity* concerns whether or not the results of the study generalize beyond the specific features of the study. Sometimes, students think they have identified a confound when, in fact, external validity is at issue. For example, suppose you read about a problem-solving study using anagrams in which the independent variable is problem difficulty (two levels—easy and hard) and the dependent variable is the number of problems solved in 15 minutes. You learn the subjects in the study are students in the college's honors program. Thinking critically, you believe the study is flawed and would be better if it tested students with a wider range of ability—you might say the study is confounded by using only honors students. You might be right about who should be in the study, but you would be wrong to call it confounding. You have confused internal and external validity. There would only be a confound in this case if honors students were in one group (e.g., the easy problems) and non-honors students were in the other group (e.g., the hard problems). Then, any differences could be due to the type of problem encountered or the type of student doing the problem solving—a classic confound. Criticizing the use of honors students is therefore not a problem with internal validity but a problem with external validity; the results might not generalize to other types of students. Keep this distinction in mind as you work through Applications Exercise 5.2; we have deliberately put in some external validity issues, in addition to the confounds you will be identifying.

Self Test 5.3

1. Determined to get into graduate school, Jan takes the GRE nine times. In her first 7 attempts, she 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. What is the best way to control for the effects of history, maturation, and regression?
3. How can attrition produce an effect similar to a subject selection effect?

This concludes our introduction to the experimental method. The next three chapters will elaborate. In Chapter 6, we distinguish between-subjects designs from within-subjects (or repeated measures) designs, and we also describe a number of control problems in experimental research. In particular, we explore the problems of creating equivalent groups in between-subjects designs, controlling for order 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 describe 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 are the experimental conditions or comparisons 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, and failure to control them leads to a problem called confounding. When a confound exists, the results could be due to the independent variable or the confounding variable. Dependent variables are the behaviors measured in the study; they must be defined precisely (operationally).

Subject Variables

Some research in psychology compares groups of participants who differ from each other in some way before the study begins (e.g., gender, age, shyness). When this occurs, the independent variable of interest 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 (e.g., Bandura's Bobo doll study). In a well-controlled study, conclusions about cause and effect can be drawn for manipulated variables but not for subject variables.

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 and precisely (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. 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. What was Robert Woodworth's definition of an experiment in psychology?
2. With anxiety as an example, illustrate the difference between independent variables that are (a) manipulated variables and (b) subject variables.
3. Distinguish between Mill's methods of Agreement and Difference, and apply them to a study with an experimental and a control group.
4. Use examples to show the differences between situational, task, and instructional independent variables.
5. What is a confound and why does the presence of one make it impossible to interpret the results of a study?
6. When a study uses subject variables, it is said that causal conclusions cannot be drawn. Why not?
7. Describe the circumstances that could reduce the statistical conclusion validity of an experiment.

8. Describe the three types of circumstances in which external validity can be reduced.
9. Explain why using speed dating is a good illustration of ecological validity.
10. Explain how the presence of a control group can help reduce threats to internal validity. Use history, maturation, and regression to the mean as specific examples.
11. As threats to internal validity, distinguish between testing and instrumentation.
12. Use Brady et al.'s (1958) study of "ulcers in executive monkeys" to illustrate subject selection effects.
13. What is attrition, when is it likely to occur, and why is it a problem?
14. Explain how internal and external validity are sometimes confused when students critically examine a research report.

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 non-manipulated subject variables. For the manipulated variables, indicate whether they are situational, task, or instructional 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. Half the students are told to visually imagine the campus buildings before they point; the remaining students are not given any specific strategy to use. Participants are asked to indicate (on a scale of 1 to 10) how confident they are about their pointing accuracy; 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 at all. 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 dogs (i.e., 12 hours without food) and not-so-hungry dogs (i.e., 6 hours without food) 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 told 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 given an elevated maze to learn (no side walls); others have to learn a more traditional alley maze (with side walls). The maze pattern is the same in both cases. Half the rats in each condition are wild rats; the remaining rats were bred in the laboratory. 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 Wal-Mart. Nearby, an experimenter records whether or not people provide any help.
7. In a memory study, a researcher wishes to know how well people can recall the locations of items in an environment. Girls and boys (ages 8–10) are compared. Each is shown a sheet of paper containing line drawings of 30 objects. For half the subjects, the items on the sheet are stereotypically male-oriented (e.g., a football); the remaining subjects get stereotypically female-oriented items (e.g., a measuring cup). After studying the objects on the first sheet for 3 minutes, all subjects are then shown a second sheet in which some of the items have moved to a new location on the page. Subjects are told to circle the objects that have moved to a new location.
8. In a study of cell phone use and driving, some participants try to perform as accurately as they can in a driving simulator (i.e., keep the car on a narrow road) while talking on a hand-held cell phone, others while talking on a

hands-free phone, and yet others without talking on a phone at all. Half the subjects have 2 years of driving experience. The remaining participants have 4 years of driving experience.

Exercise 5.2. Spot the Confound(s)

For each of the following, identify the independent and dependent variables and 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). Be sure not to confuse internal and external validity. 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 patients with schizophrenia to judge time durations. It is hypothesized that loud noise will adversely affect their judgment. Participants are tested two ways. In the "quiet" condition, some participants are tested in a small soundproof room 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 in older adults if they use visual imagery. Participants (all women over the age of 65) 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 "computer-phobia," rote participants also sit at the computer terminal, but the computer is programmed to read the lists to them. After the word lists have been presented, participants have a minute 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 (Felix) is generally well dressed, but the second (Oscar) 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 Felix or shabbily-dressed Oscar. All of the testing sessions occur between 8 and 9 o'clock in the evening, with Felix working on Monday and Oscar 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 (one dependent variable per item).
- **c.** Create operational definitions for your independent and dependent variables.

1. People are 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 when an audience is watching.
4. Animals learn a difficult maze best when they are moderately aroused. They do poorly in difficult mazes when their arousal is high or low. When the maze is easy, performance improves steadily from low to moderate to high arousal.
5. Caffeine improves memory, but only for older people.
6. In a bratwurst eating contest, those scoring high on a "sensation-seeking" scale will consume more, and this is especially true for fans of the Pittsburgh Steelers, compared with Baltimore Ravens fans.

Answers to 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 must 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 at whether self-reported frustration levels differed for those given impossible problems to solve, compared to others 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 confounds.
   External → results generalize beyond the confines of the study.
4. Ecological.

✓5.3

1. Somewhere around 1075 to 1100; regression to the mean.
2. Add a control group.
3. 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.