PART THREE

Experimental Methods

CHAPTER SEVEN

Independent Groups Designs

CHAPTER OUTLINE

Overview

In Chapter 2 we introduced you to the four goals of research in psychology: description, prediction, explanation, and application. Psychologists use observational methods to develop detailed descriptions of behavior, often in natural settings. Survey research methods allow psychologists to describe people's attitudes and opinions. Psychologists are able to make predictions about behavior and mental processes when they discover measures and observations that covary (correlations). Description and prediction are essential to the scientific study of behavior, but they are not sufficient for understanding the causes of behavior. Psychologists also seek explanation—the "why" of behavior. We achieve scientific explanation when we identify the causes of a phenomenon. Chapters 7, 8, and 9 focus on the best available research method for identifying causal relationships—*the experimental method*. We will explore how the experimental method is used to test psychological theories as well as to answer questions of practical importance.

As we have emphasized, the best overall approach to research is the *multimethod approach*. We can be more confident in our conclusions when we obtain comparable answers to a research question after using different methods. Our conclusions are then said to have *convergent validity*. Each method has different shortcomings, but the methods have complementary strengths that overcome these shortcomings. The special strength of the experimental method is that it is especially effective for establishing cause-and-effect relationships. In this chapter we discuss the reasons researchers conduct experiments and we examine the underlying logic of experimental research. Our focus is on a commonly used experimental design—the random groups design. We describe the procedures for forming random groups and the threats to interpretation that apply specifically to the random groups design. Then we describe the procedures researchers use to analyze and interpret the results they obtain in experiments, and also explore how researchers establish the external validity of experimental findings. We conclude the chapter with consideration of two additional designs involving independent groups: the matched groups design and the natural groups design.

Why Psychologists Conduct Experiments

- Researchers conduct experiments to test hypotheses about the causes of behavior.
- Experiments allow researchers to decide whether a treatment or program effectively changes behavior.

One of the primary reasons that psychologists conduct experiments is to make empirical tests of hypotheses they derive from psychological theories. For example, Pennebaker (1989) developed a theory that keeping in thoughts and feelings about painful experiences might take a physical toll. According to this "inhibition theory," it's physically stressful to keep these experiences to oneself.

Pennebaker and his colleagues conducted many experiments in which they assigned one group of participants to write about personal emotional events and another group to write about superficial topics. Consistent with the hypotheses derived from the inhibition theory, participants who wrote about emotional topics had better health outcomes than participants who wrote about superficial topics. Not all the results, however, were consistent with the inhibition theory. For example, students asked to dance expressively about an emotional experience did not experience the same health benefits as students who danced and wrote about their experience. Pennebaker and Francis (1996) did a further test of the theory and demonstrated that cognitive changes that occur through writing about emotional experiences were critical in accounting for the positive health outcomes.

Our brief description of testing the inhibition theory illustrates the general process involved when psychologists do experiments to test a hypothesis derived from a theory. If the results of the experiment are consistent with what is predicted by the hypothesis, then the theory receives support. On the other hand, if the results differ from what was expected, then the theory may need to be modified and a new hypothesis developed and tested in another experiment. Testing hypotheses and revising theories based on the outcomes of experiments can sometimes be a long and painstaking process, much like combining the pieces to a puzzle to form a complete picture. The self-correcting interplay between experiments and proposed explanations is a fundamental tool psychologists use to understand the causes of the ways we think, feel, and behave.

Well-conducted experiments also help to solve society's problems by providing vital information about the effectiveness of treatments in a wide variety of areas. This role of experiments has a long history in the field of medicine (Thomas, 1992). For example, near the beginning of the 19th century, typhoid fever and delirium tremens were often fatal. The standard medical practice at that time was to treat these two conditions by bleeding, purging, and other similar "therapies." In an experiment to test the effectiveness of these treatments, researchers randomly assigned one group to receive the standard treatment (bleeding, purging, etc.) and a second group to receive nothing but bed rest, good nutrition, and close observation. Thomas (1992) describes the results of this experiment as "unequivocal and appalling" (p. 9): The group given the standard medical treatment of the time did worse than the group left untreated. Treating such conditions using early-19th-century practices was worse than not treating them at all! Experiments such as these contributed to the insight that many medical conditions are self-limited: The illness runs its course, and patients recover on their own.

Logic of Experimental Research

- Researchers manipulate an independent variable in an experiment to observe the effect on behavior, as assessed by the dependent variable.
- Experimental control allows researchers to make the causal inference that the independent variable *caused* the observed changes in the dependent variable.
- Control is the essential ingredient of experiments; experimental control is gained through manipulation, holding conditions constant, and balancing.

- An experiment has internal validity when it fulfills the three conditions required for causal inference: covariation, time-order relationship, and elimination of plausible alternative causes.
- When confounding occurs, a plausible alternative explanation for the observed covariation exists, and therefore, the experiment lacks internal validity. Plausible alternative explanations are ruled out by holding conditions constant and balancing.

A true experiment involves the *manipulation* of one or more factors and the *measurement* (observation) of the effects of this manipulation on behavior. As you saw in Chapter 2, the factors the researcher controls or manipulates are called the *independent variables*. An independent variable must have at least two levels (also called conditions). One level may be considered the "treatment" condition and a second level is called the control (or comparison) condition. Often, more than two levels are used for additional comparisons between groups. The measures used to observe the effect (if any) of the independent variables are called *dependent variables*. One way to remember the distinction between these two types of variables is to understand that the outcome (dependent variable) *depends* on the independent variable.

Experiments are effective for testing hypotheses because they allow us to exercise a relatively high degree of control in a situation. Researchers use control in experiments to be able to state with confidence that the independent variable *caused* the observed changes in the dependent variable. The three conditions needed to make a causal inference are covariation, time-order relationship, and elimination of plausible alternative causes (see Chapter 2).

Covariation is met when we observe a relationship between the independent and dependent variables of an experiment. A time-order relationship is established when researchers manipulate an independent variable and *then* observe a subsequent difference in behavior (i.e., the difference in behavior is contingent on the manipulation). Finally, elimination of plausible alternative causes is accomplished through the use of control procedures, most importantly, through *holding conditions constant* and *balancing*. When the three conditions for a causal inference are met, the experiment is said to have **internal validity,** and we can say the independent variable *caused* the difference in behavior as measured by the dependent variable. Let us first describe a research situation in which these conditions are *not* met. Then we'll describe a published experiment in which they are met; that is, the experiment can be said to have internal validity.

Key Concept

The three conditions for causal inference are not met in a kind of research study called a *one-group pretest-posttest design* (Campbell & Stanley, 1966). This design typically involves one group of participants who are singled out for treatment or intervention. Observations of behavior are made before (pretest) and after (posttest) the treatment. Such would be the case, for example, if children in one classroom were instructed using a new way of teaching mathematics (treatment) with relevant math tests given before and after the new method is introduced. This design can be described as

Stage 1	Stage 2	Stage 3
O_1	X	O_2

where O_1 refers to the first observation of the group, or pretest; X indicates a treatment; and O_2 refers to the second observation, or posttest.

Although this particular research design is sometimes used in psychological research, the design has very little internal validity, and it is difficult to interpret any results from this type of research design. For example, should a difference between pretest and posttest measures be found, it's possible this difference could be due to some event *other than* the treatment, or the group of participants simply changed over time naturally. Consider again the case in which one classroom is introduced to a new way to teach math and tests are given before and after the treatment. We don't know whether the children also may have been helped by parents or tutors during this period, nor do we know how much children might simply be improving at math as they mature cognitively. This particular design has so little going for it in terms of allowing cause-and-effect inferences that it is sometimes referred to as a "pre-experimental design," or one that serves as a "bad experiment" to illustrate possible threats to internal validity (see Campbell & Stanley, 1966). We will have much more to say about threats to internal validity later in this chapter and especially in Chapter 11 when we discuss research in natural settings. However, let us now examine a different research design, one that does have internal validity.

Random Groups Design

- In an independent groups design, each group of subjects participates in only one condition of the independent variable.
- Random assignment to conditions is used to form comparable groups by balancing or averaging subject characteristics (individual differences) across the conditions of the independent variable manipulation.
- When random assignment is used to form independent groups for the levels of the independent variable, the experiment is called a random groups design.

Key Concept

In an **independent groups design,** each group of subjects participates in a different condition of the independent variable.[1] The logic of the design is straightforward. The groups are formed so as to be similar on all important characteristics at the start of the experiment. Next, in the experiment itself, the groups are treated the same except for the level of the independent variable. Thus, any difference between the groups on the dependent variable must be caused by the independent variable.

An Example of a Random Groups Design

The logic of the experimental method and the application of control techniques that produce internal validity can be illustrated in an experiment investigating girls' dissatisfaction with their body, conducted in the United Kingdom by

[1] Another term for independent groups design is *between-subjects design*. Both terms are used to describe studies in which groups of participants are compared and there is no overlap of participants in the groups of the study (i.e., each participant is in only one condition).

FIGURE 7.1 In the United States, 99% of young girls aged 3–10 have at least one Barbie, and the typical young girl has eight Barbie dolls (Rogers, 1999).

Dittmar, Halliwell, and Ive (2006). Their goal was to determine whether exposure to very thin body images causes young girls to experience negative feelings about their own body. Many experiments conducted with adolescent and adult participants demonstrate that women report greater dissatisfaction about themselves after exposure to a thin female model compared to other types of images. Dittmar and her colleagues sought to determine whether similar effects are observed for girls as young as 5 years old. The very thin body image they tested was the Barbie doll. Anthropological studies that compare the body proportions of Barbie to actual women reveal that the Barbie doll has very unrealistic body proportions, yet Barbie has become a sociocultural ideal for female beauty.

In the experiment small groups of young girls (5½–6½ years old) were read a story about "Mira" as she went shopping for clothes and prepared to go to a birthday party. As they heard the story, the girls looked at picture books with six scenes related to the story. In one condition of the experiment, the picture books had images of Barbie in the scenes of the story (e.g., shopping for a party outfit,

FIGURE 7.2 The “Emme” doll was introduced in 2002 to promote a more realistic body image for young girls. The doll is based on the U.S. supermodel named Emme.

getting ready for the party). In a second condition the picture books had similar scenes but the figure pictured was the “Emme” doll. The Emme fashion doll is an attractive doll with more realistic body proportions, representing a U.S. dress size 16. Finally, in the third condition of the experiment the picture books did not depict Barbie or Emme (or any body) but, instead, showed neutral images related to the story (e.g., windows of clothes shops, colorful balloons). These three versions of the picture books represent three levels of the independent variable that was manipulated in the experiment. Because different groups of girls participated in each level of the independent variable, the experiment is described as an independent groups design.

Manipulation Dittmar et al. (2006) used the control technique of *manipulation* to test their hypotheses about girls’ body dissatisfaction. The three conditions of the independent variable allowed these researchers to make comparisons relevant to their hypotheses. If they tested only the Barbie condition, it would be impossible to determine whether those images influenced girls’ body dissatisfaction in any way. Thus, the neutral-image condition created a comparison—a way to see if the girls’ body dissatisfaction differed depending on whether they looked at a thin ideal *vs.* neutral images. The Emme condition added an important comparison also. It is possible that *any* images of bodies might influence girls’ perceptions of themselves. Dittmar and her colleagues tested the hypothesis that only thin body ideals, as represented by Barbie, would cause body dissatisfaction.

At the end of the story, the young girls turned in their picture books and completed a questionnaire designed for their age level. Although Dittmar and her colleagues used a number of measures designed to assess the girls' satisfaction with their body, we will focus on one measure, the Child Figure Rating Scale. This scale has two rows of seven line drawings of girls' body shapes ranging from very thin to very overweight. The girls were asked first to color in the figure in the top row that most looks like her own body right now (a measure of perceived actual body shape). Then, on a second row of the same figures, each girl was asked to color in the figure that shows the way she most wants to look (ideal body shape). Girls were told they could pick any of the figures and that they could choose the same figure in each row. A body shape dissatisfaction score, the dependent variable, was computed by counting the number of figures between each girl's actual shape and her ideal shape. A score of zero indicates no body shape dissatisfaction, a negative score indicates a desire to be thinner, and a positive score indicates a desire to be bigger.

The results of this experiment were clear: Young girls exposed to the images of Barbie were more dissatisfied with their body shape than were girls who were exposed to the Emme images or to the neutral images. The average body-dissatisfaction score for the 20 girls in the Emme condition and for the 20 girls in the neutral-image condition was zero. In contrast, the average dissatisfaction score for the 17 girls in the Barbie-image condition was −.76, indicating their desire to be thinner. Through the control technique of manipulation, the first two requirements for causal inference were met in this experiment: (1) Differences in the girls' body dissatisfaction covaried with the conditions of the experiment and (2) body dissatisfaction came after viewing the images (time-order relationship). The third requirement for causal inference, elimination of alternative explanations, was accomplished in this experiment through holding conditions constant and balancing.

Holding Conditions Constant In Dittmar et al.'s experiment, several factors that could have affected the girls' attitudes toward their body were kept the same across the three conditions. All of the girls heard the same story about shopping and attending a birthday party, and they looked at their picture books for the same amount of time. They all received the same instructions throughout the experiment and received the exact same questionnaire at the conclusion. Researchers use *holding conditions constant* to make sure that the independent variable is the *only* factor that differs systematically across the groups.

If the three groups had differed on a factor other than the picture books, then the results of the experiment would have been uninterpretable. Suppose the participants in the Barbie condition had heard a different story, for example, a story about Barbie being thin and popular. We wouldn't know whether the observed difference in the girls' body dissatisfaction was due to viewing the images of Barbie or to the different story. When the independent variable of interest and a different, potential independent variable are allowed to covary, a *confounding* is present. When there are no confoundings, an experiment has *internal validity*.

Holding conditions constant is a control technique that researchers use to avoid confoundings. By holding constant the story the girls heard in the three conditions, Dittmar and her colleagues avoided confoundings by this factor. In general, a factor that is held constant cannot possibly covary with the manipulated independent variable. More importantly, a factor that is held constant does not change, so it cannot possibly covary with the dependent variable either. Thus, researchers can rule out factors that are held constant as potential causes for the observed results.

It is important to recognize, however, that we choose to control only those factors we think might influence the behavior we are studying—what we consider *plausible* alternative causes. For instance, Dittmar et al. held constant the story the girls heard in each condition. It is unlikely, however, that they controlled factors such as the room temperature to be constant across the conditions because room temperature probably would not likely affect body image (at least when varying only a few degrees). Nevertheless, we should constantly remain alert to the possibility that there may be confounding factors in our experiments whose influence we had not anticipated or considered.

Key Concept

Key Concept

Balancing Clearly, one key to the logic of the experimental method is forming comparable (similar) groups at the start of the experiment. The participants in each group should be comparable in terms of various characteristics such as their personality, intelligence, and so forth (also known as *individual differences*). The control technique of *balancing* is required because these factors often cannot be held constant. **Random assignment** of subjects to conditions is used to form comparable groups prior to implementing the independent variable. The goal of random assignment is to establish equivalent groups of participants by balancing, or averaging, individual differences across the conditions. When random assignment to conditions is used, the independent groups design is called a **random groups design.** The random groups design may be described as follows:

Stage 1	Stage 2	Stage 3
R_1	X_1	O_1
R_2	X_2	O_1
R_3	X_3	O_1

where R_1, R_2, and R_3 refer to the random assignment of subjects to the three independent conditions of the experiment; X_1 is one level of an independent variable (e.g., Barbie), X_2 is a second level of the independent variable (e.g., Emme), and X_3 is a third level of the independent variable (e.g., neutral images). An observation of behavior (O_1) in each group is then made. Unlike the one-group pretest-posttest design, the random groups design is an example of a good experiment.

In the Dittmar et al. (2006) study of girls' body image, if participants viewing the Barbie images were shown to be more overweight or to own more Barbie dolls than participants viewing the Emme or neutral images, a plausible alternative explanation for the findings exists. It's possible that being overweight or having more Barbie dolls, not the version of the images, could explain why participants in the Barbie condition experienced greater body dissatisfaction. (In the language of the researcher, a confounding would be present.) Similarly, individual differences in the girls' body dissatisfaction *before* the experiment was conducted could be a reasonable alternative explanation for the study's findings. When random assignment is used to balance these individual differences across the groups, however, we rule out the alternative explanation that any differences we obtain between the groups on the dependent variable are due to characteristics of the participants.

When we balance a factor such as body weight, we make the three groups equivalent in terms of their *average* body weight. Note that this differs from holding body weight constant, which would require that all of the girls in the study have the same body weight. Similarly, balancing the number of Barbie dolls owned by girls in the three groups would mean that the *average* number of dolls owned in the three groups is the same, not that the number of dolls owned by each girl is held constant at some number. The beauty of random assignment is that *all* individual differences are balanced, not just the ones we've mentioned. Therefore, we can rule out alternative explanations due to *any* individual differences among participants.

In summary, Dittmar and her colleagues concluded that exposure to thin body images, such as Barbie, *causes* young girls to be dissatisfied with their own bodies. They were able to make this conclusion because they

- manipulated an independent variable that varied the images girls viewed,
- eliminated other plausible explanations through holding relevant conditions constant, and
- balanced individual differences among the groups through random assignment to conditions.

Box 7.1 summarizes how Dittmar and her colleagues applied the experimental method, specifically, the random groups design, to their study of young girls' body image.

BOX 7.1

SUMMARY OF GIRLS' BODY IMAGE EXPERIMENT

Overview of experimental procedure. Young girls (ages 5½–6½) were assigned to look at one of three different picture books while listening to a story. After viewing the books, participants answered questions about their body image.

Independent variable. Version of picture book viewed by participants. (Barbie, Emme, or neutral images).

Dependent variable. Body dissatisfaction measured by assessing the difference between girls' actual body image and their ideal body image.

Explanation of control procedures

Holding conditions constant. Girls in the three conditions listened to the same story, were given the same instructions, and answered the same questions at the conclusion.

Balancing. Individual differences among the girls were balanced through random assignment to different experimental conditions.

Explanation of experimental logic providing evidence for causality

Covariation. The girls' body dissatisfaction was found to vary with experimental condition.

Time-order relationship. The version of the picture book was manipulated prior to measuring body dissatisfaction.

Elimination of plausible alternative causes. Control procedures of holding conditions constant and balancing individual differences through random assignment protected against confoundings.

Conclusion. Exposure to very thin body images (the Barbie picture books) caused body dissatisfaction.

(Based on Dittmar, Halliwell, & Ive, 2006)

STRETCHING EXERCISE

In this exercise you are to respond to the questions that appear after this brief description of an experiment.

Bushman (2005) examined whether people's memory for advertisements is affected by the type of television program they watch. Participants (N = 336, ages 18–54) were randomly assigned to watch one of four types of television programs: violent (e.g., *Cops*), sexually explicit (e.g., *Sex and the City*), violence and sex (e.g., *CSI Miami*), or neutral (e.g., *America's Funniest Animals*). Within each TV program were embedded the same 12 (30-second) ads. To make sure participants were likely to have equal exposure to the brands represented in the ads, the researchers selected relatively unfamiliar brands (e.g., "Dermoplast," "José Olé"). Three commercial breaks, each with four ads, were placed at approximately 12, 24, and 36 minutes into each program. Two random orders of ads were used. Participants were tested in small groups, and each session was conducted in a comfortable setting in which participants sat in padded chairs and were provided soft drinks and snacks. After they watched the program, participants received surprise memory tests for the content of the ads. The results indicated that memory for the advertised brands was poorer when the television program contained violence or sex. Memory impairment for ads was greatest for programs that contained sexually explicit material.

1. What aspect of the experiment did Bushman (2005) control by using manipulation?
2. What aspect of the experiment did Bushman control by holding conditions constant?
3. What aspect of the experiment did Bushman control by using balancing?

From Bushman, B. J. (2005). Violence and sex in television programs do not sell products in advertisements. *Psychological Science, 16,* 702–708.

Block Randomization

- Block randomization balances subject characteristics and potential confoundings that occur during the time in which the experiment is conducted, and it creates groups of equal size.

Key Concept

A common procedure for carrying out random assignment is **block randomization.** First, let us describe exactly how block randomization is carried out, and then we will look at what it accomplishes. Suppose we have an experiment with five conditions (labeled, for convenience, as A, B, C, D, and E). One "block" is made up of a random order of all five conditions:

One block of conditions	→	Random order of conditions
A B C D E		C A E B D

In block randomization, we assign subjects to conditions one block at a time. In our example with five conditions, five subjects would be needed to complete the first block with one subject in each condition. The next five subjects would be assigned to one of each of the five conditions to complete a second block, and so on. If we want to have 10 subjects in each of five conditions, then there would be 10 blocks in the block-randomized schedule. Each block would consist of a random arrangement of the five conditions. This procedure is illustrated below for the first 11 participants.

10 Blocks	*Participants*		*Condition*	
1) C A E B D	1) Cara	→	C	First block
2) E C D A B	2) Andy	→	A	
3) D B E A C	3) Jacob	→	E	
4) B A C E D	4) Molly	→	B	
5) A C E D B	5) Emily	→	D	
6) A D E B C	6) Eric	→	E	Second block
7) B C A D E	7) Anna	→	C	
8) D C A E B	8) Laura	→	D	
9) E D B C A	9) Sarah	→	A	
10) C E B D A	10) Lisa	→	B	
	11) Tom	→	D	

and so on for 50 participants

There are several advantages when block randomization is used to randomly assign subjects to groups. First, block randomization produces groups that are of equal size. This is important because the number of observations in each group affects the reliability of the descriptive statistics for each group, and it is desirable to have the reliability of these measures comparable across groups. Block randomization accomplishes this. Second, block randomization controls for time-related variables. Because experiments often take a substantial amount of time to complete, some participants can be affected by events that occur during the time the experiment is conducted. In block randomization, every condition is tested in each block so these time-related variables are balanced across the conditions of the experiment. If, for example, a traumatic event occurs on a campus in which an experiment is being conducted, the number of participants who experienced the event will be equivalent in each condition if block randomization is used. We assume, then, that the effects of the event on participants' performance will be equivalent across the conditions. Block randomization also works to balance other time-related variables, such as changes in experimenters or even changes in the populations from which subjects are drawn. For example, a perfectly acceptable experiment could be done drawing students from both fall and spring semester classes if a block randomization schedule is used. The beauty of block randomization is that it will balance (or average) any characteristics of participants (including the effects of time-related factors) across the conditions of an experiment.

If you want to practice the procedure of block randomization, you can do Challenge Question 1A at the end of this chapter.

Threats to Internal Validity

- Randomly assigning intact groups to different conditions of the independent variable creates a potential confounding due to pre-existing differences among participants in the intact groups.
- Block randomization increases internal validity by balancing extraneous variables across conditions of the independent variable.
- Selective subject loss, but not mechanical subject loss, threatens the internal validity of an experiment.

- Placebo control groups are used to control for the problem of demand characteristics, and double-blind experiments control both demand characteristics and experimenter effects.

We've seen that *internal validity* is the degree to which differences in performance on a dependent variable can be attributed clearly and unambiguously to an effect of an independent variable, as opposed to some other uncontrolled variable. These uncontrolled variables are often referred to as **threats to internal validity.** These threats are potential alternative explanations for a study's findings. For example, when we discussed the one-group pretest-posttest design, we described how an event other than the treatment might cause a difference between pre- and post-performance. In order to make a clear cause-and-effect inference about an independent variable, threats to internal validity must be controlled. We next describe problems in experimental research that can result in threats to internal validity, and methods to control these threats.

Key Concept

Testing Intact Groups Random assignment is used to form comparable groups in the random groups design. There are times, however, when *non*comparable groups are formed even when random assignment appears to have been used. This problem occurs when intact groups (not individuals) are randomly assigned to the conditions of an experiment. Intact groups are formed prior to the start of the experiment. For example, the different sections of an introductory psychology course are intact groups. Students are not randomly assigned to different sections of introductory psychology (although sometimes scheduling classes seems random!). Students often choose to be in a particular section because of the time the class meets, the instructor, friends who will be in the class, and any number of other factors. If a researcher were to randomly assign different sections to levels of an independent variable, a confounding due to testing intact groups could occur.

The source of the confounding due to noncomparable groups arises when individuals differ systematically across the intact groups. For example, students who choose to take an 8 A.M. section may differ from students who prefer an 11 A.M. section. Random assignment of these intact groups to experimental conditions is simply not sufficient to balance the systematic differences among the intact groups. These systematic differences between the two intact groups are almost guaranteed to threaten the internal validity of the experiment. The solution to this problem is simple—do not use intact groups in a random groups design.

Balancing Extraneous Variables A number of factors in an experiment may vary as a result of practical considerations when carrying out the study. For example, to complete an experiment more quickly, a researcher might decide to have several different experimenters test small groups of participants. The sizes of the groups and the experimenters themselves become potentially relevant variables that could confound the experiment. For example, if all the individuals in the experimental group were tested by one experimenter and all of those in the control group were tested by another experimenter, the levels of the intended independent variable would become confounded with the two experimenters. We would not be able to determine whether an observed difference between the two groups was due to the independent variable or to the fact that different experimenters tested participants in the experimental and control groups.

Potential variables that are not directly of interest to the researcher but that could still be sources of confounding in the experiment are called *extraneous variables*. But don't let the term fool you! An experiment confounded by an extraneous variable is no less confounded than if the confounding variable were of considerable inherent interest. For example, Evans and Donnerstein (1974) found that students who volunteer for research participation early in an academic term are more academically oriented and are more likely to have an internal locus of control (i.e., they emphasize their own responsibility, rather than external factors, for their actions) than students who volunteer late in a term. Their findings suggest it would not be wise to test all of the participants in the experimental condition at the beginning of the term and participants in the control condition at the end of the term, as this would potentially confound the independent variable with characteristics of the participants (e.g., locus of control, academic focus).

Block randomization controls extraneous variables by balancing them across groups. All that is required is that entire blocks be tested at each level of the extraneous variable. For example, if there were four different experimenters, entire blocks of the block-randomized schedule would be assigned to each experimenter. Because each block contains all the conditions of the experiment, this strategy guarantees that each condition will be tested by each experimenter. Usually, we would assign the same number of blocks to each experimenter, but this is not essential. What is essential is that entire blocks be tested at each level of the extraneous variable, which, in this case, is the four experimenters. The balancing act can become a bit tricky when there are several extraneous variables, but careful advance planning can avoid confounding by such factors.

Subject Loss We have emphasized that the logic of the random groups design requires that the groups in an experiment differ only because of the levels of the independent variable. We have seen that forming comparable groups of subjects at the beginning of an experiment is another essential characteristic of the random groups design. It is equally important that the groups be comparable except for the independent variable at the end of the experiment. When subjects begin an experiment but fail to complete it successfully, the internal validity of the experiment can be threatened. It is important to distinguish between two ways in which subjects can fail to complete an experiment: mechanical subject loss and selective subject loss.

Key Concept

Mechanical subject loss occurs when a subject fails to complete the experiment because of an equipment failure (in this case, the experimenter is considered part of the equipment). Mechanical subject loss can occur if a computer crashes, or if the experimenter reads the wrong set of instructions, or if someone inadvertently interrupts an experimental session. Mechanical loss is a less critical problem than selective subject loss because the loss is not related to any characteristic of the subject. As such, mechanical loss should not lead to systematic differences between the characteristics of the subjects who successfully complete the experiment in the different conditions of the experiment. Mechanical loss can also reasonably be understood as the result of chance events that should occur equally across groups. Hence, internal validity is not typically threatened when subjects must be excluded from the experiment due to mechanical loss. When mechanical subject loss occurs, it should be documented. The name of the dropped subject and the reason for the loss should be recorded. The lost subject can then be replaced by the next subject tested.

Key Concept

Selective subject loss is a far more serious matter. **Selective subject loss** occurs (1) when subjects are lost differentially across the conditions of the experiment; (2) when some characteristic of the subject is responsible for the loss; and (3) when this subject characteristic is related to the dependent variable used to assess the outcome of the study. Selective subject loss destroys the comparable groups that are essential to the logic of the random groups design and can thus render the experiment uninterpretable.

We can illustrate the problems associated with selective subject loss by considering a fictitious but realistic example. Assume the directors of a fitness center decide to test the effectiveness of a 1-month fitness program. Eighty people volunteer for the experiment, and they randomly assign 40 to each of two groups. Random assignment to conditions creates comparable groups at the start of the experiment by balancing individuals' characteristics such as weight, fitness level, motivation, and so on across the two groups. Members of the control group are simply asked to take a fitness test at the end of the month. Those in the experimental group participate in a vigorous fitness program for 1 month prior to the test. Assume all 40 control participants show up for the fitness test at the end of the month, but only 25 of the experimental participants stay with the rigorous fitness program for the full month. Also assume that the average fitness score for the 25 people remaining in the experimental group is significantly higher than the average score for the 40 people in the control group. The directors of the fitness center then make the claim, "A scientifically based research study has shown that our program leads to better fitness."

Do you think the fitness center's claim is justified? It's not. This hypothetical study represents a classic example of selective subject loss, so the results of the study can't be used to support the fitness center's claim. The loss occurred differentially across conditions; participants were lost only from the experimental group. The problem with differential loss is not that the groups ended up different in size. The results would have been interpretable if 25 people had been randomly assigned to the experimental group and 40 to the control group and all the individuals had completed the experiment. Rather, selective subject loss is a problem because the 25 experimental participants who completed the fitness program are not likely to be comparable to the 40 control participants. The 15 experimental participants who could not complete the rigorous program are likely to have been less fit (even before the program began) than the 25 experimental participants who completed the program. The selective loss of participants in the experimental group likely destroyed the comparable groups that were formed by random assignment at the beginning of the experiment. In fact, the final fitness scores of the 25 experimental participants might have been higher than the average in the control group even if they had not participated in

FIGURE 7.3 Many people who begin a rigorous exercise program fail to complete it. In a sense, only the "fittest" survive, a situation that could cause problems of interpretation if different types of fitness programs were being compared.

the fitness program because they were more fit when they began! Thus, the subject loss in this experiment meets the other two conditions for selective subject loss. Namely, the loss is likely due to a characteristic of the participants—their original level of fitness—and this characteristic is relevant to the outcome of the study.

If selective subject loss is not identified until after the experiment is completed, little can be done except to chalk up the experience of having conducted an uninterpretable experiment. Preventive steps can be taken, however, when researchers realize in advance that selective loss may be a problem. One alternative is to administer a pretest and screen out subjects who are likely to be lost. For example, in the exercise study, an initial test of fitness could have been given, and only those participants who scored above some minimal level would have participated in the experiment. Screening participants in this way would involve a potential cost. The results of the study would likely apply only for people above the minimal fitness level. This cost may be well worth paying because an interpretable study of limited generality is still preferable to an uninterpretable study.

There is a second preventive approach that researchers can use when facing the possibility of selective subject loss. Researchers can give all subjects a pretest but then simply randomly assign participants to conditions. Then, if a subject is lost from the experimental group, a subject with a comparable pretest score can be dropped from the control group. In a sense, this approach tries to restore the initial comparability of the groups. Researchers must be able to anticipate possible factors that could lead to selective subject loss, and they must make sure their pretest measures these factors.

Key Concept

Placebo Control and Double-Blind Experiments The final challenge to internal validity we will describe arises because of expectations held by both participants and experimenters. Demand characteristics represent one possible source of bias due to participants' expectations (Orne, 1962). *Demand characteristics* refer to the cues and other information that participants use to guide their behavior in a psychological study (see Chapter 4). For example, research participants who know they have been given alcohol may expect to experience certain effects, such as relaxation or giddiness. They may then behave consistent with these expectations rather than in response to the effects of the alcohol per se. Potential biases can also arise due to the expectations of the experimenters. The general term used to describe these biases is **experimenter effects** (Rosenthal, 1963, 1994a). Experimenter effects may be a source of confounding if experimenters treat subjects differently in the different groups of the experiment in ways other than those required to implement the independent variable. In an experiment involving alcohol, for instance, experimenter effects could occur if the experimenters read the instructions more slowly to participants who had been given alcohol than to those who had not. Experimenter effects also can occur when experimenters make biased observations based on the treatment a subject has received. For example, biased observations might arise in the alcohol study if the experimenters were more likely to notice unusual motor movements or slurred speech among the "drinkers" (because they "expect" drinkers to behave this way). (See discussion of expectancy effects in Chapter 4.)

Key Concept

Researchers can never eliminate the problems of demand characteristics and experimenter effects, but there are special research designs that control these problems. Researchers use a **placebo control group** as one way to control demand characteristics. A *placebo* (from the Latin word meaning "I shall please") is a substance that looks like a drug or other active substance but is actually an inert, or inactive, substance. Some research even indicates that there can be therapeutic effects from the placebo itself (e.g., Kirsch & Sapirstein, 1998), based on participants' expectations for an effect of a "drug." Researchers test the effectiveness of a proposed treatment by comparing it to a placebo. Both groups have the same "awareness" of taking a drug and, therefore, similar expectations for a therapeutic effect. That is, the demand characteristics are similar for the groups—participants in both groups expected to experience effects of a drug. Any differences between the experimental groups and the placebo control group could legitimately be attributed to the actual effect of the drug taken by the experimental participants, and not the participants' expectations about receiving a drug.

Key Concept

The use of placebo control groups in combination with a double-blind procedure can control for both demand characteristics and experimenter effects. In a **double-blind procedure,** both the participant and the observer are blind to (unaware of) what treatment is being administered. In an experiment testing the effectiveness of a drug treatment, two researchers would be needed to accomplish the double-blind procedure. The first researcher would prepare the drug capsules and code each capsule in some way; the second researcher would distribute the drugs to the participants, recording the code for each drug as it was given to an individual. This procedure ensures there is a record of which drug each person received, but neither the participant nor the experimenter who actually administers the drugs (and observes their effects) knows which treatment the participant received. Thus, experimenter expectancies about the effects of the treatment are controlled because the researcher who makes the observations is unaware of who received the treatment and who received the placebo. Similarly, demand characteristics are controlled because participants remain unaware of whether they received the drug or placebo.

Experiments that involve placebo control groups are a valuable research tool for assessing the effectiveness of a treatment while controlling for demand characteristics. The use of placebo control groups, however, does raise special ethical concerns. The benefits of the knowledge gained using placebo control groups must be evaluated in light of the risks involved when research participants who expect to receive a drug may instead receive a placebo. Typically, the ethics of this procedure are addressed in the informed consent procedure prior to the start of the experiment. Participants are told they may receive a drug or a placebo. Only individuals who consent to receiving either the placebo or the drug participate in the research. Should the experimental drug prove effective, then the researchers are ethically required to offer the treatment to participants in the placebo condition.

Analysis and Interpretation of Experimental Findings

The Role of Data Analysis in Experiments

- Data analysis and statistics play a critical role in researchers' ability to make the claim that an independent variable has had an effect on behavior.
- The best way to determine whether the findings of an experiment are reliable is to do a replication of the experiment.

A good experiment, as is true of all good research, begins with a good research question. We have described how researchers use control techniques to design and implement an experiment that will allow them to gather interpretable evidence to answer their research question. This part of the research process is similar to what detectives do in a criminal investigation. Detectives carefully gather evidence to determine if the person they suspect is, in fact, the one who committed the crime. The most thorough investigation, however, is not sufficient to "make the case" that the suspect is guilty. Prosecuting attorneys must present the evidence to a jury, and their case must be compelling enough to withstand the counterarguments presented by defense attorneys. Similarly, researchers cannot "make their case" by simply conducting a good experiment. They must also present the evidence in a convincing way to demonstrate that their findings support their conclusions based on that evidence. Data analysis and statistics play a critical role in the analysis and interpretation of experimental findings.

Robert Abelson, in his book *Statistics as Principled Argument* (1995), suggests that the primary goal of data analysis is to determine whether our observations support a claim about behavior. That is, can we "make the case" for our conclusion based on the evidence we have gathered in our experiment? We provide a more complete description of how researchers use data analysis and statistics in Chapters 12 and 13. Here we will introduce the central concepts of data analysis that apply to the interpretation of the results of experiments. But first let us mention one very important way that researchers can make their case concerning the results of their research.

Key Concept

The best way to determine whether the findings obtained in an experiment are reliable (consistent) is to replicate the experiment and see if the same outcome is obtained. **Replication** means repeating the procedures used in a particular experiment in order to determine whether the same results will be obtained a second time. As you might imagine, an exact replication is almost impossible to carry out. The subjects tested in the replication will be different from those tested in the original study; the testing rooms and experimenters also may be different. Nevertheless, replication is still the best way to determine whether a research finding is reliable. If we required, however, that the reliability of every experiment be established by replication, the process would be cumbersome and inefficient. Participants for experiments are a scarce resource, and doing a replication means we won't be doing an experiment to ask new and different questions about behavior. Data analysis and statistics provide researchers with an alternative to replication for determining whether the results of a single experiment can be used to make a claim about the effect an independent variable has on behavior.

Stat Tip

Data analysis of an experiment involves three stages: (1) getting to know the data, (2) summarizing the data, and (3) confirming what the data reveal. In the first stage we try to find out what is going on in the data set, look for errors, and make sure the data make sense. In the second stage we use descriptive statistics and graphical displays to summarize what was found. In the third stage we seek evidence for what the data tell us about behavior. In this stage we make our conclusions about the data using various statistical techniques.

In the following sections we provide only a brief introduction to these stages of data analysis. A more complete introduction to data analysis is found in Chapters 12 and 13 (see especially Box l2.1). These later chapters will become particularly important if you need to read and interpret the results of a psychology experiment published in a scientific journal or if you carry out your own psychology experiment.

We will illustrate the process of data analysis by examining the results of an experiment that investigated the effects of rewards and punishments while participants played violent video games. Carnagey and Anderson (2005) noted that a large body of research evidence demonstrates that playing violent video games increases aggressive affect, cognitions, and behavior. They wondered, however, whether the effects of violent video games would differ when players are *punished* for violent game actions compared to when the same actions are *rewarded* (as in most video games). In previous research with televised violence, participants who witnessed punishment of violence were less aggressive than participants who watched violence that was not punished. Thus, one hypothesis formed by Carnagey and Anderson was that when violent video-game actions are punished, players would be less aggressive. Another hypothesis, however, stated that when punished for their violent actions, players would become frustrated and therefore more aggressive.

In Carnagey and Anderson's studies, undergraduate participants played one of three versions of the same competitive race-car video game ("Carmageddon 2") in a laboratory setting. In the reward condition, participants were rewarded (gained points) for killing pedestrians and race opponents (this is the unaltered version of the game). In the punishment condition, the video game was altered so that participants lost points for killing or hitting opponents. In a third condition, the game was altered to be nonviolent and participants gained points for passing checkpoints as they raced around the track (all pedestrians were removed and race opponents were programmed to be passive).

Carnagey and Anderson (2005) reported the results of three experiments in which participants were randomly assigned to play one of the three versions of the video game. The primary dependent variables were participants' hostile emotions (Experiment 1), aggressive thinking (Experiment 2), and aggressive behaviors (Experiment 3). Across the three studies, participants who were rewarded for violent actions in the video game were higher in aggressive emotions, cognitions, and behavior compared to the punishment and nonviolent game conditions. Punishing aggressive actions in the video game caused participants to experience greater hostile emotions (similar to the reward condition) relative to nonviolent play, but did not cause them to experience increased aggressive cognitions and behavior.

In order to illustrate the process of data analysis, we will examine more closely Carnagey and Anderson's results for aggressive cognitions (Experiment 2). After playing one of the three video-game versions, participants completed a word fragment task in which they were asked to complete as many words (out of 98) as they could in 5 minutes. Half of the word fragments had aggressive possibilities. For example, the word fragment "K I ___ ___" could be completed as "kiss" or "kill" (or other possibilities). Aggressive cognition was operationally defined as the proportion of word fragments a participant completed with aggressive words. For example, if a participant completed 60 of the word fragments in 5 minutes and 12 of those expressed aggressive content, the participant's aggressive cognition score would be .20 (i.e., 12/60 = .20).

Describing the Results

- The two most common descriptive statistics that are used to summarize the results of experiments are the mean and standard deviation.
- Measures of effect size indicate the strength of the relationship between the independent and dependent variables, and they are not affected by sample size.

- One commonly used measure of effect size, *d*, examines the difference between two group means relative to the average variability in the experiment.
- Meta-analysis uses measures of effect size to summarize the results of many experiments investigating the same independent variable or dependent variable.

Data analysis should begin with a careful inspection of the data set with special attention given to possible errors or anomalous data points. The next step is to describe what was found. At this stage the researcher wants to know "What happened in the experiment?" To begin to answer this question, researchers use *descriptive statistics*. The two most commonly reported descriptive statistics are the mean (a measure of central tendency) and the standard deviation (a measure of variability). The means and standard deviations for aggressive cognition in the video-game experiment are presented in Table 7.1. The means show that aggressive cognition was highest in the reward condition (.210) and lowest in the nonviolent condition (.157). Aggressive cognition in the punishment condition (.175) fell between the nonviolent and reward conditions. We can note that for participants in the reward condition, approximately 1 in 5 words was completed with aggressive content (remember, though, that only half of the word fragments had aggressive possibilities).

In a properly conducted experiment, the standard deviation in each group should reflect only individual differences among the subjects who were randomly assigned to that group. Subjects in each group should be treated in the same way, and the level of the independent variable to which they've been assigned should be implemented in the same way for each subject in the group. The standard deviations shown in Table 7.1 indicate that there was variation around the mean in each group and that the variation was about the same in all three groups.

Key Concept

One important question researchers ask when describing the results of an experiment is how large an effect the independent variable had on the dependent variable. Measures of **effect size** can be used to answer this question because they indicate the strength of the relationship between the independent and dependent variables. One advantage of measures of effect size is that they are not influenced by the size of the samples tested in the experiment. Measures of effect size take into account more than the mean difference between two conditions in an experiment. The mean difference between two groups is always

TABLE 7.1 MEAN AGGRESSIVE COGNITION, STANDARD DEVIATIONS, AND CONFIDENCE INTERVALS FOR THE THREE CONDITIONS OF THE VIDEO-GAME EXPERIMENT

Video-game version	Mean	*SD*	.95 Confidence interval*
Reward	.210	.066	.186–.234
Punishment	.175	.046	.151–.199
Nonviolent	.157	.050	.133–.181

*Confidence intervals were estimated based on data reported in Carnagey and Anderson (2005).

Key Concept

relative to the average variability in participants' scores. One frequently used measure of effect size is **Cohen's *d*.** Cohen (1992) developed procedures that are now widely accepted. He suggested that *d* values of .20, .50, and .80 represent small, medium, and large effects of the independent variable, respectively.

We can illustrate the use of Cohen's *d* as a measure of effect size by comparing two conditions in the video-game experiment, the reward condition and the nonviolent condition. The *d* value is .83 based on the difference between the mean aggressive cognition in the reward condition (.210) and the nonviolent condition (.157). This *d* value allows us to say that the video-game independent variable, reward *vs*. nonviolent, had a large effect on the aggressive cognition in these two conditions. Effect-size measures provide researchers with valuable information for describing the findings of an experiment.

Stat Tip

Measures of central tendency and variability, as well as effect size, are described in Chapters 12 and 13. In those chapters we outline the computational steps for these measures and discuss their interpretation. Many different effect-size measures are found in the psychology literature. In addition to Cohen's *d*, for example, a popular measure of effect magnitude is eta squared, which is a measure of the strength of association between the independent and dependent variables (see Chapter 13). That is, eta squared estimates the proportion of total variance accounted for by the effect of the independent variable on the dependent variable. Measures of effect magnitude are most helpful when comparing the numeric values of a measure from two or more studies or when averaging measures across studies as is done when a meta-analysis is performed (see below).

Key Concept

Researchers also use measures of effect size in a procedure called meta-analysis. **Meta-analysis** is a statistical technique used to summarize the effect sizes from several independent experiments investigating the same independent or dependent variable. Meta-analyses are used to answer questions like: Are there gender differences in conformity? What are the effects of class size on academic achievement? Is cognitive therapy effective in the treatment of depression? Box 7.2 describes a meta-analysis of studies on effective psychotherapy for youth with psychological disorders. The results of individual experiments, no matter how well done, often are not sufficient to provide answers to questions about such important general issues. We need to consider a body of literature (i.e., many experiments) pertaining to each issue. (See Hunt, 1997, for a good and readable introduction to meta-analysis.)

Meta-analysis allows us to draw stronger conclusions about the principles of psychology because these conclusions emerge only after looking at the results of many individual experiments. Each single strand contributes to the strength of a rope, but the rope is stronger than any strand. Similarly, each properly done experiment strengthens our confidence in a particular psychological principle. The results of any individual experiment represent a strand in the stronger principles of psychology.

BOX 7.2

AN EXAMPLE OF META-ANALYSIS: "EVIDENCE-BASED YOUTH PSYCHOTHERAPIES VERSUS USUAL CLINICAL CARE"

Weisz, Jensen-Doss, and Hawley (2006) used meta-analysis to summarize the results of 32 psychotherapy studies with youth that compared the effects of "evidence-based treatments" and "usual care." An evidence-based treatment (EBT) is one that has received empirical support—that is, it has been shown in clinical research to help individuals. Although it seems obvious that EBTs should be widely used in clinical practice because of this empirical support, many therapists argue that these treatments would not be effective in usual clinical contexts. EBTs are structured and require therapists to follow a treatment manual. Some clinicians argue that EBTs are inflexible, rigid treatments that cannot be individualized according to clients' needs. Furthermore, opponents of EBTs argue that empirical studies that indicate effectiveness typically involve clients with less severe or less complicated problems than those seen in usual clinical practice. These arguments suggest that usual care (UC) in the form of psychotherapy, counseling, or case management as regularly conducted by mental health providers would better meet the needs of the clients typically seen in community settings.

Weisz and his colleagues used meta-analysis to compare directly the outcomes associated with EBTs and usual care. Across 32 studies that compared EBT and UC, the average effect size was 0.30. Thus, youth treated with an evidence-based treatment were better off, on average, than youth treated with usual care. The value of 0.30 falls between Cohen's (1988) criteria for small and medium effects. This effect size represents the difference between the two types of treatments, not the effect of psychotherapy per se. Weisz et al. note that when EBTs are contrasted with no-treatment control groups (e.g., waiting list), the effect sizes for EBT typically range from 0.50 to 0.80 (medium-to-large effects). In additional analyses the authors grouped studies according to factors such as the severity and complexity of treated problems, treatment settings, and characteristics of the therapists. These analyses were done to determine whether the concerns voiced by critics of evidence-based treatments warrant the continued use of usual care. Weisz and his colleagues found that grouping studies according to these various factors did not influence the overall outcome that EBTs outperformed UC.

This meta-analysis allows psychologists to make the claim with more confidence for a general psychological principle regarding psychotherapy: Evidence-based treatments provide better outcomes for youth than usual care.

Meta-analyses provide an efficient and effective way to summarize the results of large numbers of experiments using effect-size measures. Nevertheless, the sophisticated statistical techniques that are used in meta-analyses are powerful only when the data from the studies being analyzed have been gathered in appropriate ways. The results of meta-analyses can be misleading when experiments with poor internal validity are included. Thus, important questions regarding meta-analyses ask: Which experiments should be included in the meta-analysis? Will only experiments reported in journals with high editorial standards be included, or will the meta-analysis include research reports that have not undergone editorial review? In general, the methodological quality of the experiments included in the meta-analysis will determine its ultimate value (see Judd, Smith, & Kidder, 1991).

Confirming What the Results Reveal

- Researchers use inferential statistics to determine whether an independent variable has a reliable effect on a dependent variable.
- Two methods to make inferences based on sample data are null hypothesis testing and confidence intervals.
- Researchers use null hypothesis testing to determine whether mean differences among groups in an experiment are greater than the differences that are expected simply because of error variation.
- A statistically significant outcome is one that has a small likelihood of occurring if the null hypothesis were true.
- Researchers determine whether an independent variable has had an effect on behavior by examining whether the confidence intervals for different samples in an experiment overlap. The degree of overlap provides information as to whether the sample means estimate the same population mean or different population means.

Perhaps the most basic claim that researchers want to make when they do an experiment is that the independent variable did have an effect on the dependent variable. Another way to phrase this claim is to say that researchers want to confirm that the independent variable *produced a difference in behavior*. Descriptive statistics alone are not sufficient evidence to confirm this basic claim.

To confirm whether the independent variable has produced an effect in an experiment, researchers use *inferential statistics*. They need to use inferential statistics because of the nature of the control provided through random assignment in experiments. As we have previously described, random assignment does not *eliminate* the individual differences among subjects. Random assignment simply *balances,* or averages, the individual differences among subjects (comparably) across the groups of the experiment. The nonsystematic (i.e., random) variation due to differences among subjects within each group is called *error variation*. The presence of error variation poses a potential problem because the means of the different groups in the experiment may differ simply because of error variation, not because the independent variable has an effect. Thus, by themselves, the mean results of the best-controlled experiment do not permit a definite conclusion about whether the independent variable has produced a difference in behavior. Inferential statistics allow researchers to test whether differences between group means are due to an effect of the independent variable, not just due to chance (error variation). Researchers use two types of inferential statistics to make decisions about whether an independent variable has had an effect: null hypothesis testing and confidence intervals.

Stat Tip

We realize that it may be frustrating to learn that the results of the best-controlled experiment often do not permit a definite conclusion about whether the independent variable produced a difference in behavior. In other words, what you have learned so far about research methods is not enough! Unfortunately, even with the tools of data analysis we cannot give you a way to make *definite* conclusions about what produced a difference in behavior. But what we can give you is a way (actually, several ways) to make the best possible statement about what produced a difference. The conclusion will be based on a *probability*—namely, a probability that will help you to decide whether your effect is or is not simply due to chance. It is easy to get lost in the complexities of null hypothesis testing and confidence intervals, but keep in mind the following two critical points:

First and foremost, differences in behavior can arise simply due to chance (often referred to as *error variation*). What you want to know is, how likely it is that the difference you have observed is only due to chance (not to the effect of your independent variable)? Actually, what you would really like to know is, how likely it is that your independent variable had an effect. However, we can't answer these questions using statistical inference. As you will see, statistical inference is indirect (see, for example, Box 13.1 in Chapter 13).

Second, the data you have collected represent *samples* from a population; but in a sense, it is *populations*, not samples, that really matter. (If only sample means mattered, then you could simply look at the sample means to see if they were different.) The mean performance for the samples in the various conditions of your experiment provides estimates that are used to *infer* the mean of the population. When you make statements of statistical inference, you are using the sample means to make decisions (inferences) about differences between (or among) population means. Once again we refer you to Chapter 13 for a more complete discussion of these issues.

Key Concept

Key Concept

Null Hypothesis Significance Testing (NHST) Researchers most frequently use **null hypothesis significance testing (NHST)** to decide whether an independent variable has produced an effect in an experiment. Null hypothesis significance testing begins with the assumption that the independent variable has had *no* effect. If we assume that the null hypothesis is true, we can use probability theory to determine the probability that the difference we did observe in our experiment would occur "by chance." *A **statistically significant** outcome is one that has only a small likelihood of occurring if the null hypothesis were true*. A statistically significant outcome means only that the difference we obtained in our experiment is larger than would be expected if error variation alone (i.e., chance) were responsible for the outcome.

The outcome of an experiment is usually expressed in terms of the differences between the means for the conditions in the experiment. How do we know the probability of the obtained outcome in an experiment? Most often, researchers use inferential statistics tests such as the *t*-test or *F*-test. The *t*-test is used when there are two levels of the independent variable, and the *F*-test is used when there are three or more levels of the independent variable. Each value of a *t*- or *F*-test has a probability value associated with it when the null hypothesis is assumed. This probability can be determined once the researcher has computed the value of the test statistic.

Just how small does the probability of our outcome need to be in order to be statistically significant? Scientists tend to agree that outcomes with probabilities (p) of less than 5 times out of 100 (or $p < .05$) are judged to be statistically significant. The probability value researchers use to decide that an outcome is statistically significant is called the *level of significance*. The level of significance is indicated by the Greek letter alpha (α).

We can now illustrate the procedures of null hypothesis testing to analyze the video-game experiment we described earlier (see Table 7.1). The first research question we would ask is whether there was any *overall* effect of the independent variable of video-game version. That is, did aggressive cognition differ as a function of the three versions of the video game? The null hypothesis for this overall test is that there is no difference among the population means represented by the means of the experimental conditions (remember that the null hypothesis assumes no effect of the independent variable). The p value for the F-test that was computed for the effect of the video-game version was less than the .05 level of significance; thus, the overall effect of the video-game variable was statistically significant. To interpret this outcome, we would need to refer to the descriptive statistics for this experiment in Table 7.1. There we see that the mean aggressive cognition for the three video-game conditions was different. For example, aggressive cognition was highest with the reward video game (.210) and lowest with the nonviolent video game (.157). The statistically significant outcome of the F-test allows us to make the claim that the video-game version did produce a difference in aggressive cognition.

Researchers want to make more specific claims about the effects of independent variables on behavior than that the independent variable did have an effect. F-tests of the overall differences among the means tell us that something happened in the experiment, but they don't tell us much about what did happen. To gain this more specific information about the effects of independent variables, researchers can use confidence intervals.

Key Concept

Using Confidence Intervals to Examine Mean Differences The confidence intervals for each of the three groups in the video-game experiment are shown in Table 7.1. A confidence interval is associated with a probability (usually .95) that the interval contains the true population mean. The width of the interval tells us how precise our estimate is (the narrower the better). **Confidence intervals** can also be used to compare differences between two population means. We can use the .95 confidence intervals presented in Table 7.1 to ask specific questions about the effects of the video-game version on aggressive cognition. We accomplish this by examining whether the confidence intervals for the different video-game groups overlap. *When the confidence intervals do not overlap, we can be confident that the population means for the two groups differ*. For example, notice that the confidence interval for the reward group is .186 to .234. This indicates that there is a .95 probability that the population mean for aggressive cognition in the reward condition falls between .186 and .234 (remember the sample mean of .210 only *estimates* the population mean). The confidence interval for the nonviolent group is .133 to .181. This confidence interval does not overlap at all with the confidence interval for the reward group (i.e., the upper limit of .181 for the nonviolent group is less than the lower limit of .186 for the reward group). With this evidence we can make the claim that aggressive cognition in the reward condition was greater than aggressive cognition in the nonviolent video-game condition.

When we compare the confidence intervals for the reward group (.186–.234) and the punishment group (.151–.199), however, we come to a different conclusion. The confidence intervals for these groups do overlap. Even though the sample means of .210 and .175 differ, we cannot conclude that the population means differ because of the overlap of the confidence intervals. We can offer the following rule of thumb for interpreting this result: *If intervals overlap slightly, then we must acknowledge our uncertainty about the true mean difference and postpone judgment; if the intervals overlap such that the mean of one group lies within the interval of another group, we may conclude that the population means* ***do not*** *differ*. In the video-game experiment, the overlap is small and the sample means for each condition do not fall within the intervals for the other group. We want to decide whether the populations differ, but all we can really say is that we don't have sufficient evidence to decide one way or the other. In this situation we must postpone judgment until the next experiment is done.

The logic and computational procedures for confidence intervals and for the *t*-test are found in Chapter 12. The *F*-test (in its various forms) is discussed in Chapter 13. Confidence intervals and rules for their interpretation are found in Chapter 12 (see especially Box 12.5).

What Data Analysis Can't Tell Us

We've already alluded to one thing that our data analysis can't tell us. Even if our experiment is internally valid and the results are statistically significant, we cannot say *for sure* that our independent variable had an effect (or did not have an effect). We must learn to live with probability statements. The results of our data analysis also can't tell us whether the results of our study have practical value or even if they are meaningful. It is easy to do experiments that ask trivial research questions (see Sternberg, 1997, and Chapter 1). It is also easy (maybe too easy!) to do a bad experiment. Bad experiments—that is, ones that lack internal validity—can easily produce statistically significant outcomes and nonoverlapping confidence intervals; however, the outcome will be uninterpretable.

When an outcome is statistically significant, we conclude that the independent variable produced an effect on behavior. Yet, as we have seen, our analysis does not provide us with certainty regarding our conclusion, even though we reached the conclusion "beyond a reasonable doubt." Also, when an outcome is *not* statistically significant, we cannot conclude with certainty that the independent variable did *not* have an effect. All we can conclude is there is not sufficient evidence in the experiment to claim that the independent variable produces an effect. Determining that an independent variable has not had an effect can be even more crucial in applied research. For example, is a generic drug as effective as its brand-name counterpart? To answer this research question, researchers often seek to find no difference between the two drugs. The standards for experiments attempting to answer questions regarding no difference between conditions are higher than those for experiments seeking to confirm that an independent variable does have an effect. We describe these standards in Chapter 13.

Because researchers rely on probabilities to make decisions about the effects of independent variables, there is always some chance of making an error. There are two types of errors that can occur when researchers use inferential statistics. When we claim that an outcome is statistically significant and the null hypothesis (no difference) is really true, we are making a Type I error. A *Type I error* is like a false alarm—saying that there is a fire when there is not. When we conclude that we have insufficient evidence to reject the null hypothesis and it is, in fact, false, we are making a *Type II error* (Type I and Type II errors are described more fully in Chapter 13). We would never make either of these errors if we could know for sure whether the null hypothesis was true or false. While being mindful of the possibility that data analysis can lead to incorrect decisions, we must also remember that data analysis can and does—lead to correct decisions. The most important thing for researchers to remember is that *inferential statistics can never replace replication as the ultimate test of the reliability of an experimental outcome*.

Establishing the External Validity of Experimental Findings

- The findings of an experiment have external validity when they can be applied to other individuals, settings, and conditions beyond the scope of the specific experiment.
- In some investigations (e.g., theory-testing), researchers may choose to emphasize internal validity over external validity; other researchers may choose to increase external validity using sampling or replication.
- Conducting field experiments is one way that researchers can increase the external validity of their research in real-world settings.
- Partial replication is a useful method for establishing the external validity of research findings.
- Researchers often seek to generalize results about conceptual relationships among variables rather than specific conditions, manipulations, settings, and samples.

As you learned in Chapter 4, *external validity* refers to the extent to which findings from an experiment can be generalized to individuals, settings, and conditions beyond the scope of the specific experiment. A frequent criticism of highly controlled experiments is that they lack external validity; that is, the findings observed in a controlled laboratory experiment may describe what happens only in that specific setting, with the specific conditions that were tested, and with the specific individuals who participated. Consider again the video-game experiment in which college students played a race-car video game in a laboratory setting. The laboratory setting is ideally suited for exercising control procedures that ensure the internal validity of an experiment. But do these findings help us understand violence and aggression in a natural setting? When a different type of exposure to violence is involved? When the people exposed to violence are senior citizens? These are questions of external validity, and they raise a more general question. If the findings of laboratory experiments are so specific, what good are they to society?

One answer to this question is a bit unsettling, at least initially. Mook (1983) argued that, when the purpose of an experiment is to test a specific hypothesis derived from a psychological theory, the question of external validity of the findings is irrelevant. An experiment is often done to determine whether subjects *can* be induced to behave in a certain way. The question whether subjects *do* behave that way in their natural environment is secondary to the question raised in the experiment. The issue of the external validity of experiments is not a new one, as reflected in the following statement by Riley (1962): "In general, laboratory experiments are not set up to imitate the most typical case found in nature. Instead, they are intended to answer some specific question of interest to the experimenter" (p. 413).

Of course, researchers often do want to obtain findings that they can generalize beyond the boundaries of the experiment itself. Researchers seeking to generalize their findings can include the characteristics of the situations to which they wish to generalize in their experiments. For example, Ceci (1993) described a research program that he and his colleagues conducted on children's eyewitness testimony. He described how their research program was motivated in part because previous studies on this topic did not capture all the dimensions of an *actual* eyewitness situation. Ceci described how their research program included factors such as multiple suggestive interviews, very long retention intervals, and recollections of stressful experiences. Including these factors made the experiments more representative of situations that are actually involved when children testify.

Ceci (1993) also pointed out, however, that important differences remained between the experiments and real-life situations:

> High levels of stress, assaults to a victim's body, and the loss of control are characteristics of events that motivate forensic investigations. Although these factors are at play in some of our other studies, we will never mimic experimentally the assaultive nature of acts perpetrated on child victims, because even those studies that come closest, such as the medical studies, are socially and parentally sanctioned, unlike sexual assaults against children. (pp. 41–42)

As Ceci's comments reveal, in some situations, such as those involving eyewitness testimony about despicable acts, there may be important ethical constraints on establishing the external validity of experiments.

The external validity of research findings is frequently questioned because of the nature of the "subjects." As you are aware, many studies in psychology involve college students who participate in experiments as part of their introductory psychology course. Dawes (1991), among others, argues that college students are a select group who may not always provide a good basis for building general conclusions about human behavior and mental processes. Similarly, Sue (1999) argues that researchers' greater emphasis on internal validity over external validity lessens the attention paid to the representativeness of the people who are studied. If psychologists generally believe their findings will generalize to populations other than those specifically tested in their research,

FIGURE 7.4 How similar can experiments be to real-life situations such as children testifying in court?

there is little reason to cross-validate the findings by testing ethnic minority populations or other underrepresented populations. Questions about the external validity of research findings based on the populations being studied are especially important in applied research. In medical research, for example, effective treatments for men may not be effective for women, and effective treatments for adults may not be effective for children.

Field experiments, which we mentioned briefly in Chapter 4, are one way to increase the external validity of a research study. They can also yield practical knowledge. For example, Crusco and Wetzel (1984) investigated the effect of touching on restaurant customers. Female wait staff, working as confederates, briefly touched restaurant customers on either the hand or the shoulder when returning change. The researchers speculated that a touch on the hand would produce positive feelings toward the server. They also hypothesized that a touch on the shoulder would be seen as a sign of dominance and therefore would not be viewed positively, especially by male diners. The researchers randomly assigned 114 diners to three levels of the independent variable: *Palm Touch, Shoulder Touch,* and *No Touch* (no physical contact with the customers). The major dependent variable was the size of the tip. Both male and female diners gave a significantly larger tip after being touched briefly than when not touched. Contrary to the researchers' expectations, however, the nature of the touch did not make a difference. Both male and female diners gave equally large tips when they were touched on the hand as they did when they were touched on the shoulder. Because this experiment was carried out in a natural setting, it is more likely to be representative of "real-world" conditions. Thus, we can be more confident that the results will generalize to other real-world settings than if an artificial situation had been created in the laboratory.

The external validity of experimental findings can be established through *partial replication.* Partial replications are commonly done as a routine part of the process of investigating the conditions under which a phenomenon reliably occurs. A partial replication can help to establish external validity by showing that a similar experimental result occurs when slightly different experimental procedures are used. Consider the same basic experiment done in both a large metropolitan public university and in a small rural private college; the participants and the settings in the experiments are very different. If the same results are obtained even with these different participants and settings, we can say the findings can be generalized across these two populations and settings. Notice that neither experiment alone has external validity; it is *the findings* that occur in *both* experiments that have external validity.

Researchers can also establish the external validity of their findings by doing *conceptual replications.* What we wish to generalize from any one study are conceptual relationships among variables, not the specific conditions, manipulations, settings, or samples (see Banaji & Crowder, 1989; Mook, 1983). Anderson and Bushman (1997) provide an example illustrating the logic of a conceptual replication. Consider a study with 5-year-old children to determine if a specific insult ("pooh-pooh-head") induces anger and aggression. We could then do a replication to see if the same insult produces the same result with 35-year-old adults. As Anderson and Bushman state, the findings for 5-year-olds probably wouldn't be replicated with the 35-year-olds because " 'pooh-pooh-head' just doesn't pack the same 'punch' for 5- and 35-year-old people" (p. 21). However, if we wish to establish the external validity of the idea that "insults increase aggressive behavior," we can use different words that are meaningful insults for each population.

When Anderson and Bushman (1997) examined variables related to aggression at the conceptual level, they found that findings from experiments conducted in laboratory settings and findings from correlational studies in real-world settings were very similar. They concluded that "artificial" laboratory experiments do provide meaningful information about aggression because they demonstrate the same conceptual relationships that are observed in real-world aggression. Furthermore, laboratory experiments allow researchers to isolate the potential causes of aggression and to investigate boundary conditions for when aggression will or will not occur.

What about when results in the lab and the real world disagree? Anderson and Bushman (1997) argue that these discrepancies, rather than evidence for the weakness of either method, should be used to help us refine our theories about aggression. That is, the discrepancies should make us recognize that different psychological processes may be at work in each setting. When we increase our understanding of these discrepancies, we will increase our understanding of aggression.

Establishing the external validity of *each* finding in psychology by performing partial replications or conceptual replications would be virtually impossible. But if we take arguments like those of Dawes (1991) and Sue (1999) seriously, as indeed we should, it would appear that we are facing an impossible task. How, for instance, could we show that an experimental finding obtained with a group of college students will generalize to groups of older adults, working professionals, less educated individuals, and so forth? Underwood and Shaughnessy (1975) suggest one possible approach worth considering. Their notion is that we should assume that behavior is relatively continuous across time, subjects, and settings unless we have reason to assume otherwise. Ultimately, the external validity of research findings is likely to be established more by the good judgment of the scientific community than by definitive empirical evidence.

Matched Groups Design

- A matched groups design may be used to create comparable groups when there are too few subjects available for random assignment to work effectively.
- Matching subjects on the dependent variable task is the best approach for creating matched groups, but performance on any matching task must correlate with the dependent variable task.
- After subjects are matched on the matching task, they should then be randomly assigned to the conditions of the independent variable.

To work effectively, the random groups design requires samples of sufficient size to ensure that individual differences among subjects will be balanced through random assignment. That is, the assumption of the random groups design is that individual differences "average out" across groups. But how many subjects are required for this averaging process to work as it should? The answer is "It depends." More subjects will be needed to average out individual differences when samples are drawn from a heterogeneous population than from a homogeneous one.

We can be relatively confident that random assignment will *not* be effective in balancing the differences among subjects when small numbers of subjects are tested from heterogeneous populations. However, this is exactly the situation researchers face in several areas of psychology. For example, some developmental psychologists study newborn infants; others study the elderly. Both newborns and the elderly certainly represent diverse populations, and developmental psychologists often have available only limited numbers of participants.

One alternative that researchers have in this situation is to administer all the conditions of the experiment to all the subjects, using a repeated measures design (to be discussed in Chapter 8). Nevertheless, some independent variables require separate groups of subjects for each level. For instance, suppose researchers wish to compare two types of postnatal care for premature infants and it is not possible to give both types of care to each infant. In this situation, and many others, researchers will need to test separate groups in the experiment.

Key Concept

The matched groups design is a good alternative when neither the random groups design nor the repeated measures design can be used effectively. The logic of the **matched groups design** is simple and compelling. Instead of trusting random assignment to form comparable groups, the researcher makes the groups equivalent by matching subjects. Once comparable groups have been formed based on the matching, the logic of the matched groups design is the same as that for the random groups design. In most uses of the matched groups design, a pretest task is used to match subjects. The challenge is to select a pretest task (also called a matching task) that equates the groups on a dimension that is relevant to the outcome of the experiment. *The matched groups design is useful only when a good matching task is available.*

The most preferred matching task is one that uses the same task that will be used in the experiment itself. For example, if the dependent variable in the experiment is blood pressure, participants should be matched on blood pressure prior to the start of the experiment. The matching is accomplished by measuring the blood pressure of all participants and then forming pairs or triples or quadruples of participants (depending on the number of conditions in the experiment) who have identical or very similar blood pressures. Thus, at the start of the experiment, participants in the different groups will have, *on average,* equivalent blood pressure. Researchers can then reasonably attribute any group differences in blood pressure at the end of the study to the treatment (presuming other potential variables have been held constant or balanced).

In some experiments, the primary dependent variable cannot be used to match subjects. For example, consider an experiment that teaches participants different approaches to solving a puzzle. If a pretest were given to see how long it took individuals to solve this puzzle, the participants would likely learn the solution to the puzzle during the pretest. If so, then it would be impossible to observe differences in the speed with which different groups of participants solved the puzzle following the experimental manipulation. In this situation the next best alternative for a matching task is to use a task from the *same class or category* as the experimental task. In our problem-solving experiment, participants could be matched on their performance when solving a different puzzle from the experimental puzzle. A less preferred, but still possible, alternative for matching is to use a task that is from a *different class* than the experimental task. For our problem-solving experiment, participants could be matched on some test of general ability, such as a test of spatial ability. When using these alternatives, however, researchers must confirm that performance on the matching task correlates with the performance on the task that is used as the dependent variable. In general, as the correlation between the matching task and the dependent variable decreases, the advantage of the matched groups design, relative to the random groups design, also decreases.

Even when a good matching task is available, matching is not sufficient to form comparable groups in an experiment. For example, consider a matched groups design to compare two different methods of caring for premature infants so as to increase their body weight. Six pairs of premature infants could be matched on their initial body weight. There remain, however, potentially relevant characteristics of the participants beyond those measured by the matching task. For example, the two groups of premature infants may not be comparable in their general health or in their degree of parental attachment. It is important, therefore, to use random assignment in the matched groups design to try to balance other potential factors beyond the matching task. Specifically, after matching the infants on body weight, the pairs of infants would be randomly assigned to one of the two groups. In conclusion, *the matched groups design is a better alternative than the random groups design when a good matching task is available and when only a small number of subjects is available for an experiment that requires separate groups for each condition.*

FIGURE 7.5 Random assignment is not likely to be effective in balancing differences among subjects when small numbers of subjects from heterogeneous populations are tested (e.g., newborns). In this situation, researchers may want to consider the matched groups design.

Natural Groups Design

- Individual differences variables (or subject variables) are selected rather than manipulated to form natural groups designs.
- The natural groups design represents a type of correlational research in which researchers look for covariations between natural groups variables and dependent variables.
- Causal inferences cannot be made regarding the effects of natural groups variables because plausible alternative explanations for group differences exist.

Key Concept

Researchers in many areas of psychology are interested in independent variables that are called **individual differences variables,** or *subject variables.* An individual differences variable is a characteristic or trait that varies across individuals. Religious affiliation is an example of an individual differences variable. Researchers can't manipulate this variable by randomly assigning people to Catholic, Jewish, Muslim, Protestant, or other groups. Instead, researchers "control" the religious affiliation variable by systematically selecting individuals who *naturally* belong to these groups. Individual differences variables such as gender, introversion–extraversion, race, or age are important independent variables in many areas of psychology.

Key Concept

It is important to differentiate experiments involving independent variables whose levels are *selected* from those involving independent variables whose levels are *manipulated.* Experiments involving independent variables whose levels are selected—like individual differences variables—are called **natural groups designs.** The natural groups design is frequently used in situations in which ethical and practical constraints prevent us from directly manipulating independent variables. For example, no matter how interested we might be in the effects of major surgery on subsequent depression, we could not ethically perform major surgery on a randomly assigned group of introductory psychology students and then compare their depression symptoms with those of another group who did not receive surgery! Similarly, if we were interested in the relationship between divorce and emotional disorders, we could not randomly assign some people to get divorced. By using the natural groups design, however, we can compare people who have had surgery with those who have not. Similarly, people who have chosen to divorce can be compared with those who have chosen to stay married.

Researchers use natural groups designs to meet the first two objectives of the scientific method: description and prediction. For example, studies have shown that people who are separated or divorced are much more likely to receive psychiatric care than are those who are married, widowed, or have remained single. On the basis of studies like these, we can describe divorced and married individuals in terms of emotional disorders, and we can predict which group is more likely to experience emotional disorders.

Serious problems can arise, though, when the results of natural groups designs are used to make causal statements. For instance, the finding that divorced persons are more likely than married persons to receive psychiatric care shows that these two factors covary. This finding could be taken to mean that divorce causes emotional disorders. But, before we conclude that divorce *causes* emotional disorders, we must assure ourselves that the time-order condition for a causal inference has been met. Does divorce precede the emotional disorder, or does the emotional disorder precede the divorce? A natural groups design does not tell us.

The natural groups design also poses problems when we try to satisfy the third condition for demonstrating causality, eliminating plausible alternative causes. The individual differences studied in the natural groups design are usually confounded—groups of individuals are likely to differ in many ways *in addition* to the variable used to classify them. For example, individuals who divorce and individuals who stay married may differ with respect to a number of characteristics other than their marital status, for example, their religious practices or financial circumstances. Any differences observed between divorced and married individuals may be due to these other characteristics, not to divorce. The manipulation done by "nature" is rarely the controlled type we have come to expect in establishing the internal validity of an experiment.

There are approaches for drawing causal inferences in the natural groups design. One effective approach requires that individual differences be studied in combination with independent variables that can be manipulated. This combination of more than one independent variable in one experiment requires the use of a complex design, which we will describe in Chapter 9. For now, recognize that drawing causal inferences based on the natural groups design can be a treacherous enterprise. Although such designs are sometimes referred to as "experiments," there are important differences between an experiment involving an individual differences variable and an experiment involving a manipulated variable.

SUMMARY

Researchers conduct experiments to test hypotheses derived from theories, but experiments can also be used to test the effectiveness of treatments or programs in applied settings. The experimental method is ideally suited to identifying cause-and-effect relationships when the control techniques of manipulation, holding conditions constant, and balancing are properly implemented.

In Chapter 7 we focused on applying these control techniques in experiments in which different groups of subjects are given different treatments representing the levels of the independent variable. In the random groups design, the groups are formed using randomization procedures such that the groups are comparable at the start of the experiment. If the groups perform differently following the manipulation, it is presumed that the independent variable is responsible. Random assignment is the most common method of forming comparable groups. By distributing subjects' characteristics equally across the conditions of the experiment, random assignment is an attempt to ensure that the differences among subjects are balanced, or averaged, across groups in the experiment. The most common technique for carrying out random assignment is block randomization.

There are several threats to the internal validity of experiments that involve testing independent groups. Testing intact groups even when the groups are randomly assigned to conditions should be avoided because the use of intact groups is highly likely to result in a confounding. Extraneous variables, such as different rooms or different experimenters, must not be allowed to confound the independent variable of interest.

A more serious threat to the internal validity of the random groups design is involved when subjects fail to complete the experiment successfully. Selective subject loss occurs when subjects are lost differentially across the conditions and some characteristic of the subject that is related to the outcome of the experiment is responsible for the loss. We can help prevent selective loss by restricting subjects to those likely to complete the experiment successfully, or we can compensate for it by selectively dropping comparable subjects from the group that did not experience the loss. Demand characteristics and experimenter effects can be minimized through the use of proper experimental procedures, but they can best be controlled by using placebo control and double-blind procedures.

Data analysis and statistics provide an alternative to replication for determining whether the results of a single experiment can be used as evidence to claim that an independent variable has had an effect on behavior. Data analysis involves the use of both descriptive statistics and inferential statistics. Describing the results of an experiment typically involves the use of means, standard deviations, and measures of effect size. Meta-analysis makes use of measures of effect size to provide a quantitative summary of the results of a large number of experiments on an important research problem.

Inferential statistics are important in data analysis because researchers need a way to decide whether the obtained differences in an experiment are due to chance or are due to the effect of the independent variable. Confidence intervals and null hypothesis testing are two effective statistical techniques researchers can use to analyze experiments. Statistical analysis cannot guarantee, however, that experimental findings will be meaningful or be of practical significance. Replication remains the ultimate test of the reliability of a research finding.

Researchers also strive to establish the external validity of their experimental findings. When testing psychological theories, researchers tend to emphasize internal validity over external validity. One effective approach for establishing the external validity of findings is to select representative samples of all dimensions on which you wish to generalize. By conducting field experiments, researchers can increase the external validity of their research findings to real-world settings. Partial replications and conceptual replications are two common ways that researchers use to establish external validity.

The matched groups design is an alternative to the random groups design when only a small number of subjects is available, when a good matching task is available, and when the experiment requires separate groups for each treatment. The biggest problem with the matched groups design is that the groups are equated only on the characteristic measured by the matching task. In the natural groups design, researchers select the levels of independent variables (usually individual differences or subject variables) and look for systematic relationships between these independent variables and other aspects of behavior. Essentially, the natural groups design involves looking for correlations between subjects' characteristics and their performance. Such correlational research designs pose problems in drawing causal inferences.

Key Concepts

internal validity	**207**
independent groups designs	**208**
random assignment	**212**
random groups design	**212**
block randomization	**214**
threats to internal validity	**216**
mechanical subject loss	**217**
selective subject loss	**218**
experimenter effects	**220**
placebo control group	**220**
double-blind procedure	**220**
replication	**222**
effect size	**224**
Cohen's *d*	**225**
meta-analysis	**225**
null hypothesis significance testing (NHST)	**228**
statistically significant	**228**
confidence interval	**229**
matched groups design	**235**
individual differences variable	**237**
natural groups design	**237**

Review Questions

1 Describe two reasons why psychologists conduct experiments.

2 Describe how the control techniques of manipulation, holding conditions constant, and balancing contribute to meeting the three conditions necessary for a causal inference.

3 Explain why a research study conducted using the one-group pretest-posttest design has very little internal validity.

4 Explain why comparable groups are such an essential feature of the random groups design, and describe how researchers achieve comparable groups.

5 What preventive steps could you take if you anticipated that selective subject loss could pose a problem in your experiment?

6 Explain how placebo control and double-blind techniques can be used to control demand characteristics and experimenter effects.

7 Explain why meta-analysis allows researchers to draw stronger conclusions about the principles of psychology.

8 Explain what a statistically significant outcome of an inferential statistics test tells you about the effect of the independent variable in an experiment.

9 Explain what you could conclude if the confidence intervals did not overlap when you were testing for a difference between means for two conditions in an experiment.

10 Briefly describe four ways researchers can establish the external validity of a research finding.

11 Briefly explain the logic of the matched groups design, and identify the three conditions under which the matched groups design is a better alternative than the random groups design.

12 How do individual differences variables differ from manipulated independent variables, and why does this difference make it difficult to draw causal inferences on the basis of the natural groups design?

CHALLENGE QUESTIONS

1 An experimenter is planning to do a random groups design experiment to study the effect of the rate of presenting stimuli on people's ability to recognize the stimuli. The independent variable is the presentation rate, and it will be manipulated at four levels: Very Fast, Fast, Slow, and Very Slow. The experimenter is seeking your help and advice with the following aspects of the experiment:

A The experimenter asks you to prepare a block-randomized schedule such that there will be four participants in each of the four conditions. To do this, you can use the following random numbers that were taken from the random number table in the Appendix (Table A.1).

1-5-6-6-4-1-0-4-9-3-2-0-4-9-2-3-8-3-9-1
9-1-1-3-2-2-1-9-9-9-5-9-5-1-6-8-1-6-5-2
2-7-1-9-5-4-8-2-2-3-4-6-7-5-1-2-2-9-2-3

B The experimenter is considering restricting participants to those who pass a stringent reaction time test so as to be sure that they will be able to perform the task successfully with the Very Fast presentation rate. Explain what factors the experimenter should consider in making this decision, being sure to describe clearly what risks, if any, are taken if only this restricted set of participants is tested.

C The experimenter discovers that it will be necessary to test participants in two different rooms. How should the experimenter arrange the testing of the conditions in these two rooms so as to avoid possible confounding by this extraneous variable?

2 A researcher conducted a series of experiments on the effects of external factors that might influence people's persistence in exercise programs. In one of these experiments, the researcher manipulated three types of distraction while participants walked on a treadmill. The three types of distraction were concentrating on one's own thoughts (concentration group), listening to a tape of music (music group), and watching a video of people engaging in outdoor recreation (video group). The dependent variable was how strenuous the treadmill exercise was at the time the participant decided to end the session (the incline of the treadmill was regularly increased as the person went through the session, thereby making the exercise increasingly strenuous). In an introductory psychology course, 120 students volunteered to participate in the experiment, and the researcher randomly assigned 40 students to each of the three levels of the distraction variable. The researcher expected that the mean strenuousness score would be highest in the video group, next highest in the music group, and lowest in the concentration group.

(*continued*)

After only 2 minutes on the treadmill, each participant was given the option to stop the experiment. This brief time interval was chosen so that participants were given the option to stop before any of them could reasonably be expected to be experiencing fatigue. Data for the participants who decided to stop after only 2 minutes were not included in the analysis of the final results. Fifteen students chose to stop in the concentration group; 10 stopped in the music group; and no students stopped in the video group. The results did not support the researcher's predictions. The mean strenuousness score (on a scale from 0 to 100) for students who completed the experiment was highest for the concentration group (70), next highest for the music group (60), and lowest for the video group (50).

A Identify a possible threat to the internal validity of this experiment, and explain how this problem could account for the unexpected results of the study.

B Assume that a pretest measure was available for each of the 120 participants and that the pretest measured the degree to which each subject was likely to persist at exercise. Describe how you could use these pretest scores to confirm that the problem you identified in question 2A had occurred.

3 The newspaper headline summarizing research that had been reported in a medical journal read: "Study: Exercise Helps at Any Age." The research described in the article involved a 10-year study of nearly 10,000 men—and only men. The men were given a treadmill test between 1970 and 1989. Then they were given a second treadmill test 5 years after the first test, and their health was monitored for another 5 years. Men who were judged unfit on both treadmill tests had a death rate over the next 5 years of 122 per 10,000. Men judged fit on both treadmill tests had a 5-year death rate of only 40 per 10,000. Most interestingly, men judged unfit on the first treadmill test but fit on the second had a death rate of 68 per 10,000. The benefits of exercise were even greater when only deaths from heart attacks were examined. The benefits from exercise were present across a wide range of ages—thus, the headline.

A Why is the newspaper headline for this article potentially misleading?

B Why do you think the researchers tested only men?

C Identify two different ways of obtaining evidence that you could use to decide whether the results of this study could be applied to women. One of the ways would make use of already published research, and the other way would require doing a new study.

4 An experiment was done to test the effectiveness of a new drug that is being considered for possible use in the treatment of people who experience chronic anxiety. Fifty people who are chronically anxious are identified through a local health clinic, and all 50 people give their informed consent to participate in the experiment. Twenty-five people are randomly assigned to the experimental group, and they receive the new drug. The other 25 people are randomly assigned to the control group, and they receive the commonly used drug. The participants in both groups are monitored by a physician and a clinical psychologist during the 6-week treatment period. After the treatment period, the participants provide a self-rating on a reliable and valid 20-point scale indicating the level of anxiety they are experiencing (higher scores indicate greater anxiety). The mean self-rating in the experimental group was 10.2 ($SD = 1.5$), and the mean rating in the control group was 13.5 ($SD = 2.0$). The .95 confidence interval for the mean self-rating in the experimental group was 9.6 to 10.8. The .95 confidence interval for the control group was 12.7 to 14.3.

A Explain why a double-blind procedure would be useful in this experiment, and describe how the double-blind procedure could be carried out in this experiment.

B Focus on the descriptive statistics for this experiment. How would you describe the effect of the drug variable on anxiety ratings using the means for each condition? What do the standard deviations tell you about the anxiety ratings in the experiment?

C The probability associated with the test for the mean difference between the two groups was $p = .01$. What claim would you make about the effect of the treatment based on this probability? What claim would you make based on the estimates of the population means for the two groups in this experiment based on a comparison of the confidence intervals?

D The effect size for the treatment variable in this experiment is $d = .37$. What information does this effect size tell you about the effectiveness of the drug beyond what you know from the test of statistical significance and from comparing the confidence intervals?

Answer to Stretching Exercise

1 Bushman (2005) manipulated the independent variable of type of television program in his study. There were four levels of the independent variable: violent, sexually explicit, violent and sex, and neutral.

2 Bushman (2005) held several factors constant: the same advertisements were used in each condition, participants were tested in small groups in the same setting, and ads were placed at approximately the same point in each program.

3 Bushman (2005) balanced the characteristics of the participants across the four levels by randomly assigning participants to conditions. Thus, participants in each level were equivalent, on average, in their memory ability and their exposure to television programs and products. Bushman also used two random orders of the ads to balance any potential effects due to placement of the ads during the TV programs.

Answer to Challenge Question 1

A The first step is to assign a number from 1 to 4 to the respective conditions: 1 = Very Fast; 2 = Fast; 3 = Slow; and 4 = Very Slow. Then, using the random numbers, select four sequences of the numbers from 1 to 4. In doing this you skip any numbers greater than 4 and any number that is a repetition of a number already selected in the sequence. For example, if the first number you select is a 1, you skip all repetitions of 1 until you have selected all the numbers for the sequence of 1 to 4. Following this procedure and working across the rows of random numbers from left to right, we obtained the following four sequences for the four blocks of the randomized block schedule. The order of the conditions for each block is also presented. The block-randomized schedule specifies the order of testing the conditions for the first 16 participants in the experiment.

Block 1: 1-4-3-2 Very Fast, Very Slow, Slow, Fast

Block 2: 4-2-3-1 Very Slow, Fast, Slow, Very Fast

Block 3: 1-3-2-4 Very Fast, Slow, Fast, Very Slow

Block 4: 2-3-4-1 Fast, Slow, Very Slow, Very Fast

B The investigator is taking a reasonable step to avoid selective subject loss, but restricting participants to those who pass a stringent reaction time test entails the risk of decreased external validity of the obtained findings.

C The rooms can be balanced by assigning entire blocks from the block-randomized schedule to be tested in each room. Usually, the number of blocks assigned to each room is equal, but this is not essential. For effective balancing, however, several blocks should be tested in each room.