Designing, Conducting, Analyzing, and Interpreting Experiments with Two Groups

CHAPTER 10

Experimental Design: The Basic Building Blocks

Now that the preliminaries are out of the way, we are ready to begin an experiment. Or are we? Although we have chosen a problem, read the relevant literature, developed a hypothesis, selected our variables, instituted control procedures, and considered the participants, we are still not ready to start the experiment. Before we can actually begin, we must select a blueprint. If you were about to design a house, you would be faced with an overwhelming variety of potential plans—you would have many choices and selections ahead of you. Fortunately, selecting a blueprint for your experiment is simpler than designing a house because there are relatively few standard choices used by experimenters in designing their experiments.

Selecting a blueprint for your experiment is just as important as selecting one for a house. Can you imagine what a house would look like if you began building it without any plans? The result would be a disaster. The same is true of "building" an experiment. We refer to the research blueprint as our **experimental design**. In Chapter 1 you learned that an experimental design is the general plan for selecting participants, assigning those participants to experimental conditions, controlling extraneous variables, and gathering data. If you begin your experiment without a proper design, your experiment may "collapse" just as a house built without a blueprint might. How can an experiment collapse? We have seen students begin experiments without any direction only to end up with data that fit no known procedure for statistical analysis. We also have seen students collect data that have no

Experimental design
The general plan for selecting participants, assigning participants to experimental conditions, controlling extraneous variables, and gathering data.

bearing on their original question. Thus, we hope not only that you will use this text during your current course but also that you will keep the book and consult it as you design research projects in the future.

In this chapter we will begin developing a series of questions in a flowchart to help you select the correct design for your experiment. As Charles Brewer, distinguished professor of psychology at Furman University, is fond of saying, "If you do not know where you are going, the likelihood that you will get there borders on randomness" (Brewer, 2002, p. 503). If you don't design your experiment properly, the probability that it will answer your research question is slim. Sherlock Holmes knew this lesson well: "No, no: I never guess. It is a shocking habit—destructive to the logical faculty" (Doyle, 1927, p. 93).

When you were a child and played with Legos or Tinkertoys, you probably got a beginner's set first. This set was small and simple, but with it you learned the basics of building. As you got older, you could use larger sets that allowed you to build and create more complicated objects. The parallel between children's building sets and experimental design is striking. In both cases the beginner's set helps us learn about the processes involved so that we can use the advanced set later; the basic set forms the backbone of the more advanced set. In both cases, combining simple models increases the possibilities for building, although more complex models must still conform to the basic rules of building.

The Two-Group Design

In this chapter we examine the most basic experimental design—the two-group design—and its variations. This design is the simplest possible one that can yield a valid experiment. In research situations, we typically follow the **principle of parsimony**, also known as Occam's (or Ockham's) razor. William of Occam, a fourteenth-century philosopher, became famous for his dictum "Let us never introduce more than is required for an explanation" (McInerny, 1970, p. 370). In research, we apply the principle of parsimony to research questions, just as detectives apply the principle of parsimony to their investigations: Don't needlessly complicate the question that you are asking. The two-group design is the most parsimonious design available.

Principle of parsimony
The belief that explanations of phenomena and events should remain simple until the simple explanations are no longer valid.

Independent variable (IV)
A stimulus or aspect of the environment that the experimenter directly manipulates to determine its influences on behavior.

Dependent variable (DV)
A response or behavior that the experimenter measures. Changes in the DV should be caused by manipulation of the independent variable (IV).

How Many IVs? Figure 10-1 shows the first question we must ask in order to select the appropriate design for our experiment: "How many **independent variables (IVs)** will our experiment have?" In this chapter and the next we will deal with experimental designs that have one IV. You will remember (see Chapter 6) that an IV is a stimulus or aspect of the environment that the experimenter directly manipulates to determine its influences on behavior, which is the **dependent variable (DV)**. If you want to determine how anxiety affects test performance, for example, anxiety would be your IV. If you wish to study the effects of different therapies on depression, the different therapies would be your IV. The simplest experimental design has only one IV. We will look at research designs with more than one IV in Chapter 12.

A minority of published research studies use one IV. Does that mean that experiments with one IV are somehow poor or deficient? No, there is nothing wrong with a one-IV design; however, such a design is simple and may not yield

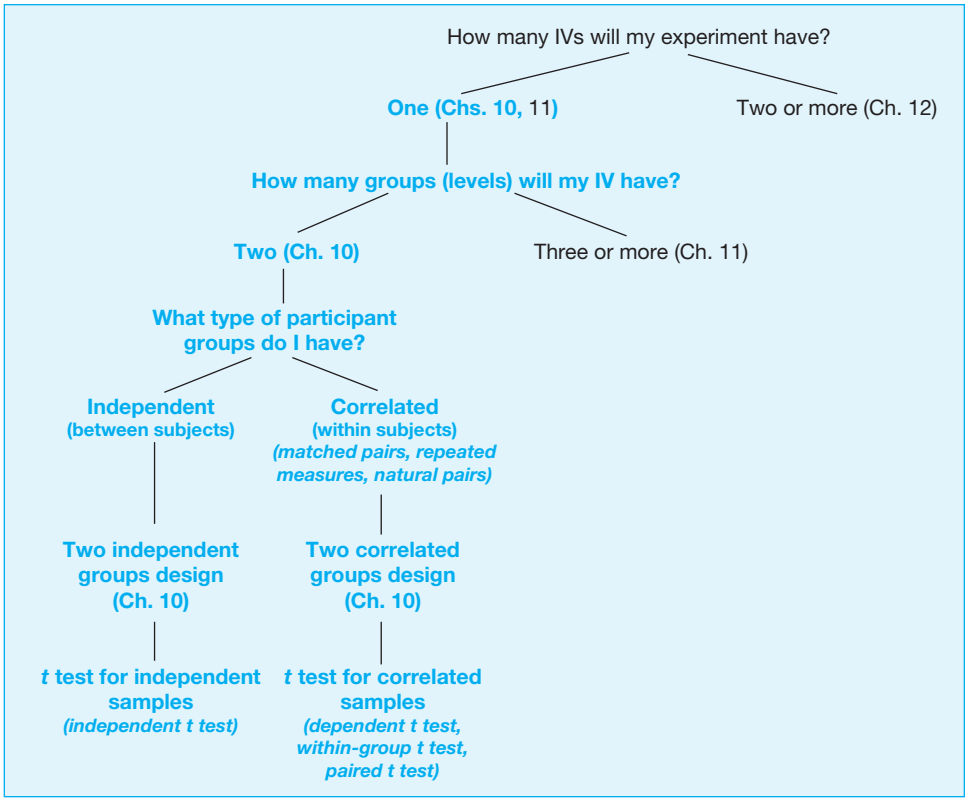


FIGURE 10-1 Experimental Design Questions.

all the answers an experimenter desires. Simple is not necessarily bad. Inexperienced researchers or researchers who are beginning to investigate new areas often prefer single-IV designs because they are easier to conduct than multiple-IV designs, and it is simpler to institute the proper control procedures. Also, the results of several one-IV experiments, when combined in one report, can describe complex phenomena.

How Many Groups? Assuming we have chosen to use a single-IV design, we come to our second question (see Figure 10-1) in determining the proper experimental design: "How many groups will I use to test my IV?" In this chapter the answer is two. Although an experiment can have a single IV, it must have at least two groups.

Why must we have two groups but only one IV?

The simplest way to determine whether our IV caused a change in behavior is to compare some research participants who have received our IV to some others who have not received the IV. If those two groups differ, and we are assured that we controlled potential **extraneous variables** (see Chapter 6), then we can conclude that the IV caused the participants to differ. The way we can test two groups with only one IV is to make the two groups differ in the amount or the type of the IV that they receive. Note carefully that the last statement is *not* the same as saying that the groups have different IVs.

Extraneous variables Uncontrolled variables that may unintentionally influence the dependent variable (DV) and thus invalidate an experiment.

The most common manner of creating two groups with one IV is to present some amount or type of IV to one group and to withhold that IV from the second group. Thus, the experimenter contrasts the *presence* of the IV with the *absence* of the IV. These differing amounts of the IV are referred to as the **levels** (also known as *treatment conditions*) of the IV. Thus, in the common two-group design, one level of the IV is none (its absence) and the other is some amount (its presence). Notice that the presence and absence of an IV is conceptualized as two differing levels of the same IV rather than as two different IVs. Now let's return to our earlier examples.

Levels Differing amounts or types of an IV used in an experiment (also known as *treatment conditions*).

If you were interested in the effects of anxiety on test performance or the effects of therapy on depression, how would you implement an experimental design so that you could compare the presence of the IV to the absence of the IV?

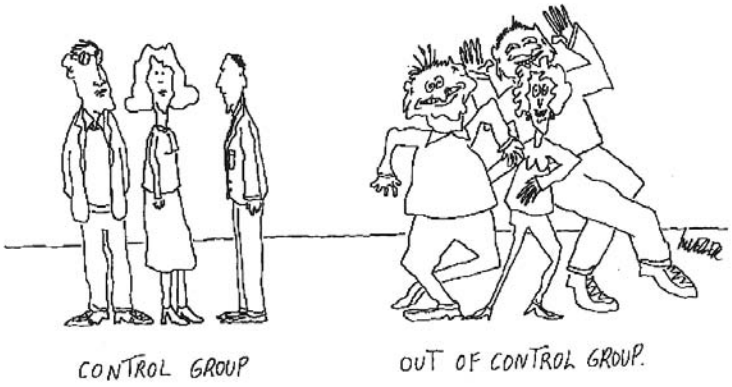
In the first example you would have to compare anxious test takers (first level) to nonanxious test takers (second level). In the second example you would compare depressed people receiving therapy (first level) to depressed people who were not receiving therapy (second level).

In this presence–absence situation, the group of participants receiving the IV is typically referred to as the **experimental group**. It is as if we were conducting the experiment on them. The group that does not receive the IV is known as the **control group**. The members of this group serve a control function because they give us an indication of how people or animals behave under "normal" conditions—that is, without being exposed to our IV. They also serve as a comparison group for the experimental group. We will use statistics to compare performance scores on the DV for the two groups to determine whether the IV has had an effect. When our two groups differ significantly, we assume that the difference is due to the IV. If the groups do not differ significantly, we conclude that our IV had no effect.

Experimental group In a two-group design, the group of participants that receives the IV.

Control group In a two-group design, the group of participants that does not receive the IV.

Let's look at a research example using the two-group design. Kristen McClellan, a student at Madonna University in Livonia, Michigan, and her faculty advisor, Edie Woods, were interested in how salesclerks reacted to customers with a disability. Their IV was the disability of hearing loss. Their experimental group of salesclerks thus encountered pairs of customers who were deaf (easily identifiable because they entered the store conversing in sign language), and the control group of clerks waited on hearing pairs of customers. McClellan and Woods (2001) randomly assigned 77 salesclerks to either the experimental or



Fortunately, we deal with control groups and experimental groups in psychological research!

control group and had the confederates (customers) enter the store. The confederates unobtrusively timed how long it took a salesperson to approach and offer assistance after initial eye contact. (They did not obtain informed consent from the clerks because the clerks were participants at minimal risk; see Chapter 2.) McClellan and Woods found that the clerks in the experimental group took significantly longer on average (3.9 minutes) to wait on the deaf customers than the clerks in the control group, who took 1.3 minutes to wait on the hearing customers. This basic two-group design is depicted in Figure 10-2. Researchers frequently use block diagrams, such as the one shown in Figure 10-2, to portray the design of an experiment graphically. Note that the IV (customer hearing) heads the entire block; the two levels of the IV comprise the two subdivisions of the block. We will use this building-block notation throughout the three chapters concerning experimental design so that you can conceptualize the various designs more easily.

Assigning Participants to Groups We have yet another question to face before we can select our experimental design: We must decide how we plan to assign our research participants to groups. (Before the 1994 edition, the American Psychological Association's *Publication Manual* referred to research participants as "subjects." You are likely to read this term in older research studies; you may also hear the term for some time before the psychological language becomes standardized—it's hard to teach old psychologists new tricks!) We can assign our participants either randomly or in some nonrandom fashion. We will examine random assignment first.

Random assignment
A method of assigning research participants to groups so that each participant has an equal chance of being in any group.

Random Assignment to Groups As you saw in Chapter 6, an often-used approach for assigning participants to groups is **random assignment**. Given that we are dealing with only two groups, we could flip a coin and assign participants to one group or the other on the basis

INDEPENDENT VARIABLE (CUSTOMER HEARING)	
EXPERIMENTAL GROUP	CONTROL GROUP
Customers were deaf	Customers were not deaf

FIGURE 10-2 The Basic Two-Group Design.

Random selection A control technique that ensures that each member of the population has an equal chance of being chosen for an experiment.

of heads or tails. As long as we flip a fair coin, our participants would have a 50-50 chance of ending up in either group. Remember that random assignment is *not* the same as **random selection**, which we also learned about in Chapter 6. Random selection deals with choosing your research participants, whereas random assignment refers to putting those participants into their groups. Random assignment is concerned with control procedures in the experiment, whereas random selection influences the generality of the results. We examined the generality issue more closely in Chapter 8.

Independent groups Groups of participants formed by random assignment.

When we randomly assign our participants to groups, we create what are known as **independent groups**. The participants in one group have *absolutely no* ties or links to the participants in the other group; they are independent of each other. If you tried to relate or pair a participant in one group to one in the other group, there would be no logical way to do so. When we wish to compare the performance of participants in these two groups, we are making what is known as a **between-subjects comparison**. We are interested in the difference *between* these two groups of participants who have no ties or links to each other.

Between-subjects comparison Refers to a contrast between groups of participants who were randomly assigned to groups.

Terminology in experimental design can sometimes be confusing—for some reason, it has never been standardized. You may hear different people refer to *independent groups designs, randomized groups designs*, or *between-subjects designs*. All these names refer to the same basic strategy of randomly assigning participants to groups. The key to avoiding confusion is to understand the principle behind the strategy of random assignment. When you understand the basic ideas, the names will make sense. As you saw in Figure 10-1, when we have one IV with two levels and participants who are assigned to groups randomly, our experiment fits the two-independent-groups design.

Random assignment is important in experimental design as a control factor. When we randomly assign our research participants to groups, we can usually assume that our two groups will now be equal on a variety of variables (Spatz, 2001). Many of these variables could be extraneous variables that might confound our experiment if left uncontrolled (see Chapter 6). Random assignment is one of those statistical procedures that is *supposed* to work—in the long run. There are no guarantees that it will create the expected outcome, however, if we select a small sample. For example, we would not be surprised to flip a coin 10 times and get 7 or 8 heads. On the other hand, if we flipped a coin 100 times, we would be quite surprised if we obtained 70 or 80 heads.

Thus, when we randomly assign participants to groups, we expect the two groups to be equal on a variety of variables that could affect the experiment's outcome. Think back to McClellan and Woods's (2001) experiment dealing with customers with and without hearing disability and salesclerks' response times. When we described their experiment, we were careful to point out that the researchers *randomly* assigned some clerks to wait on deaf customers and *randomly* assigned some clerks to wait on hearing customers. What if they had picked the most polite clerks to wait on the hearing customers (control group)? Putting the polite clerks in the control group would cause the clerks' politeness to vary systematically with levels of the IV (that is, polite clerks would wait on hearing customers, but less-polite clerks would not). Such assignment to groups would result in a **confounded experiment** (see

Confounded experiment An experiment in which an extraneous variable varies systematically with the IV, which makes drawing a cause-and-effect relation impossible.

Chapter 6). If the results showed that the control-group clerks waited on the hearing customers faster than the experimental-group clerks waited on deaf customers, we could not draw a definite conclusion about why that result happened. They might have responded more quickly because they were more polite, because they were waiting on hearing customers, or because of the combination of these two factors. Unfortunately, with a confounded experiment there is no way to determine which conclusion is appropriate. If McClellan and Woods had conducted their experiment in this manner, they would have wasted their time.

Let us remind you of one more benefit of random assignment. In Chapter 6 you learned that random assignment is the only technique we have that will help us control unknown extraneous variables. For example, in McClellan and Woods's (2001) experiment, what extraneous variables might affect salesclerks' performance? We have already identified the clerks' politeness as a possibility. The researchers did not have access to any politeness measure for the clerks. Other variables not even considered could also play a role in the clerks' performance; therefore, McClellan and Woods were careful to assign their clerks randomly to experimental and control groups. Random assignment should equate any differences between the two groups.

Can you think of a flaw in McClellan and Woods's reasoning behind random assignment?

Random assignment is a technique that *should* work in the long run. Because McClellan and Woods assigned 77 clerks to the two groups, there is a good chance that random assignment made the groups perfectly equal. If they had measured only a few salesclerks, however, random assignment might not have created equal groups. If you thought of this potential problem, congratulations! What can we do when we conduct an experiment with small numbers of participants?

Nonrandom Assignment to Groups In the previous section we saw a potential pitfall of random assignment: The groups may not be equal after all. If we begin our experiment with unequal groups, we have a problem. Remember that random assignment should create equal groups in the long run. In other words, as our groups get larger, we can place more confidence in random assignment achieving what we want it to.

Suppose we are faced with a situation in which we have few potential research participants and we are worried that random assignment may not create equal groups. What can we do? In this type of situation, we can use a nonrandom method of assigning participants to groups. What we will do is either capitalize on an existing relationship between participants or create a relationship between them. In this manner, we know something important about our participants before the experiment, and we will use **correlated assignment** (also known as *matched* or *paired assignment*) to create equal groups. Thus, we use correlated assignment whenever there is a relationship between the participants in the groups.

Correlated assignment
A method of assigning research participants to groups so that there is a relationship between small numbers of participants; these small groups are then randomly assigned to treatment conditions (also known as *paired* or *matched assignment*).

(Be careful—correlated assignment has *nothing* to do with computing a correlation coefficient.) There are three common ways to use correlated assignment.

Matched pairs
Research participants in a two-group design who are measured and equated on some variable before the experiment.

1. ***Matched Pairs.*** To create **matched pairs**, we must measure our participants on a variable (other than our IV) that could affect performance on our experiment's DV. Typically we measure a variable that could result in confounding if not controlled. After we have measured this variable, we create pairs of participants that are equal on this variable. After we have created our matched pairs, we then randomly assign participants from these pairs to the different treatment conditions.

If this description seems confusing, an example should help clarify matters. Imagine that we wanted to replicate McClellan and Woods's salesclerk study because we were worried that random assignment may not have created equal groups. Suppose we suspect that female salesclerks tend to wait on customers faster than male clerks (probably a totally fictitious supposition, but it makes a good example). In this situation we would be concerned if the ratio of female to male clerks differed between the groups. If we flip a coin to assign clerks to groups, the sex ratio of the two groups might be unequal; therefore, we decide to use matched assignment to groups. First, we create our matched pairs. The first pair consists of two clerks of the same sex; the second pair is another two clerks of the same sex. (We pair all the other clerks by sex also.) For each pair, we flip a coin to determine which clerk to assign to the experimental group and which clerk to assign to the control group. Then we repeat the procedure for our second pair, and so on. After we complete this matched assignment, we have an experimental group and a control group that are perfectly balanced in terms of sex. We have used matched assignment to create equal groups before our experiment begins. (Note: In this hypothetical example, if there were an odd number of clerks, one clerk would not have a match, and we could not use that clerk in the experiment.)

In what way is matched assignment guaranteed to create equal groups when random assignment is not?

The beauty of matched assignment is that we have measured our participants on a specific variable that could affect their performance in our experiment, and we have equated them on that variable. When we use random assignment, we are leaving this equating process to chance. Remember, we must match on a variable that could affect the outcome of our experiment. In the fictitious example we just used, you would have to be certain that sex was linked to salesclerks' performance. If you cannot measure your participants on a variable that is relevant to their performance in your experiment, then you should not use matched assignment. If you match your participants on a variable that is not relevant to their performance, then you have actually hurt your chances of finding a significant difference in your experiment (see "Statistical Issues" in the "Advantages of Correlated Group Designs" section later in this chapter).

2. ***Repeated Measures.*** In **repeated measures**, we use the ultimate in matched pairs: We simply test or measure the same participants in both treatment conditions of our experiment. The matched pairs here are perfectly equal because they consist of the same people or animals tested across the entire experiment. No extraneous variables should be able to confound this situation because any difference between the participants' performance in the two treatment conditions is due to the IV. In this type of experiment, participants serve as their own controls.

Repeated measures An experimental procedure in which research participants are tested or measured more than once.

Why is it not possible to use repeated measures in all experiments? Try to think of two reasons.

Thinking about using repeated measures for our groups forces us to consider some practical factors:

a. Can we remove the effects of the IV? McClellan and Woods (2001) did not do anything to the salesclerks that had lasting effects—waiting on deaf customers should not affect how clerks respond to other customers in the future. In a study cited in Chapter 2, Burkley et al. (2000) could not use repeated measures in their experiment because they could not remove the effects of reading a particular consent form on the students in their experiment. Think about it carefully—even though they could have allowed students to read the second form, they could not have removed the effects of reading the other form first. If the students remembered the previous form, giving them a new form would not remove its effects. Sometimes when researchers cannot remove the effects of an IV, they deal with this problem by using counterbalancing (see Chapter 6). By balancing the effects across groups, researchers hope that the effects end up equal across those groups.

b. Can we measure our DV more than once? To this point we have not focused on the DV in this chapter because it has little to do with the choice of an experimental design. When you consider using repeated measures, however, the DV is extremely important. When you use repeated measures, the DV is measured multiple times (at least twice). McClellan and Woods could have used repeated measures by having each clerk wait on both a deaf and a hearing customer. In some cases, however, it is simply not possible to use the same DV more than once. In Burkley et al.'s (2000) experiment, as soon as the students had solved the anagrams, solving the anagrams could not be used again as a DV. In other cases, we may be able to use a similar DV, but we must be cautious. To use repeated measures on solving anagrams in Burkley et al.'s experiment, we would need two different sets of anagrams, one for testing students with each type of consent form. If we use two different forms of our DV, we must ensure that they are comparable. Although we could use two sets of anagrams, they would have to be equally difficult, which

might be hard to determine. The same would be true of many different DVs such as mazes, tests, puzzles, and computer programs. If we cannot assure comparability of two measures of the same DV, then we should not use repeated-measures designs.

c. Can our participants cope with repeated testing? This question relates at least to some degree to the ethics of research, which we covered in Chapter 2. Although no specific ethical principles speak to the use of repeated measures, we should realize that requiring extended participation in our experiment might affect participants' willingness to take part in the research. Also, in extreme cases extended time in a strenuous or taxing experiment could raise concerns for physical or emotional well-being. Another of our worries in this area is whether human participants will agree to devote the amount of time that we request of them in a repeated-measures design.

It is important that we think about these practical considerations when weighing the possibility of a repeated-measures design. Although repeated-measures designs are one of our better control techniques, there are some experimental questions that simply do not allow the use of such a design.

Natural pairs Research participants in a two-group design who are naturally related in some way (e.g., a biological or social relationship).

3. ***Natural pairs.*** **Natural pairs** are essentially a combination of matched pairs and repeated measures. In this technique we create pairs of participants from naturally occurring pairs (e.g., biologically or socially related). For example, psychologists who study intelligence often use twins (natural pairs) as their research participants. This approach is similar to using the same participant more than once (repeated measures), but it allows you to compose your pairs more easily than through matching. Thus, when an experiment uses siblings, parents and children, husbands and wives, littermates, or some other biological or social relationship, that experiment has used natural pairs.

In summary, whenever there is a relationship between participants in different groups, a correlated-groups design is being used. By looking at Figure 10.1, you can see that an experiment with (a) one IV that has (b) two levels, in which you plan to use (c) correlated assignment of participants to groups, results in the *two-correlated-groups design*. Participants who have been matched on some variable or who share some relationship would have scores that are related. When we wish to compare the performance of such participants, we are making what has traditionally been known as a **within-subjects comparison**. We are essentially comparing scores within the same participants (subjects). Although this direct comparison is literally true only for repeated-measures designs, participants in matched or natural pairs are the same with regard to the matching variable.

Within-subjects comparison Refers to a contrast between groups of participants who were assigned to groups through matched pairs, natural pairs, or repeated measures.

Let's look at a student example of a two-correlated-groups design. A major area of study in psychology is the effect of stress on the body and the body's reactions. Rachel Wells (2001), a student from Nebraska Wesleyan University, found that previous researchers had used a mental arithmetic test to induce stress in participants. She wanted to determine the effects of such a test on college students' bodily reactions. She had students count backward from 715 by 13, telling them that most students could complete the task in 4 minutes. Immediately after counting for 4 minutes, she measured the participants' heart rate and blood pressure. Students then spent 10 minutes

completing questionnaires as a nonstressful rest period. After the 10-minute period, Wells measured the participants' heart rate and blood pressure again. The students showed a decrease in both heart rate and blood pressure, demonstrating that the mental arithmetic was indeed stress provoking.

Because she used repeated measures, Wells's experiment is a good example of a correlated-groups design. She measured the students' body signs after a stressful event and then again after a rest period. The measurement after the stressor was a posttest; the measurement after rest served as the comparison period. Often in such research, the experimenter might measure the body reactions *before* inducing the stressor; this measurement would be a pretest. In Wells's experiment, the IV was the stress induced, and the DVs were the students' physiological reactions to the stress. Some of the particularly important participant (subject) variables, which Wells controlled by the use of repeated measures, were participants' ages (perhaps younger or older people have different reactions to stress), time of day (perhaps physiological indicators vary by the time of day), and students' intelligence (perhaps brighter students would find the task less stressful). All these extraneous variables (and others) were controlled because the *same* students took both the comparison test and the posttest.

What if Wells had wanted to use matched pairs rather than repeated measures? Remember that matching should occur on a relevant variable—one that could be an extraneous variable if left unchecked. Suppose that these students varied widely in their math ability. Wells might have wanted to create matched pairs based on that ability, thinking that math ability would likely affect performance on the counting task, which could determine the level of stress felt by each student. Wells could have used other variables for matching as long as she was certain that the variables were related to the students' physiological indicators.

Could Wells have run this experiment using natural pairs? Based on the information given you, there is no indication that natural pairs of students existed. If the students were sets of twins, then the experiment would be ideally suited for natural pairs. It seems unlikely that there was any factor that made these students natural pairs, so if pairing was important to Wells, she would have had to create pairs by matching on some variable.

REVIEW SUMMARY

1. Psychologists plan their experiments beforehand using an **experimental design**, which serves as a blueprint for the experiment.
2. The two-group design applies to experimental situations in which one IV has two levels or conditions.
3. Two-group designs often use an **experimental group**, which receives the IV, and a **control group**, which does not receive the IV.
4. **Randomly assigning** research participants to groups results in **independent groups** of participants.
5. **Correlated groups** of research participants are formed by creating **matched pairs**, using **natural pairs**, or by measuring the participants more than once (**repeated measures**).

Check Your Progress

1. Why can't we conduct a valid experiment with only one group?
2. The differing amounts of your IV are known as the _____ of the IV.
3. How are independent groups and correlated groups different? Why is this difference important to experimental design questions?
4. Matching

1. random assignment	A. brother and sister
2. natural pairs	B. take the same test each month
3. repeated measures	C. two people with the same IQ
4. matched pairs	D. flipping a coin

5. In what type of situation do we have to be most careful when using random assignment as a control technique?
6. You are planning an experiment that could use either independent groups or correlated groups. Under what conditions should you use a correlated-groups design? When is it acceptable to use random assignment?
7. Random assignment is more likely to create equal groups when
   a. small samples are involved
   b. large samples are involved
   c. a directional hypothesis is being tested
   d. a nondirectional hypothesis is being tested

Comparing Two-Group Designs

Because there are two different two-group designs, researchers must choose whether they want to design their experiment with independent or correlated groups. You may be wondering how researchers make such a choice. In the next sections we will cover some issues that psychologists must consider when they plan their research studies. Read carefully—you may be facing this choice yourself in the future.

Look at Figure 10-1 again. You can see that the two-independent-groups design and the two-correlated-groups design are quite similar. Both designs describe experimental situations in which you use one IV with two groups. The only difference comes from how you assign your participants to groups. If you simply assign on a random basis, you use the two-independent-groups design. On the other hand, if you match your participants on some variable, if you test your participants twice, or if your participants share some relationship, you use the two-correlated-groups design.

Choosing a Two-Group Design Now that you have two experimental designs that can handle very similar experimental situations, how do you choose between them? Should you use independent groups, or should you use correlated groups of some sort?

You may remember we said that random assignment is supposed to "work" (i.e., create equal groups) in the long run. If you are using large groups of participants, therefore, random assignment should equate your groups adequately. The next question, of course, is how large is large? Unfortunately, there is no specific answer to this question—the answer may vary from researcher to researcher. If you are using 20 or more participants per group, you can

feel fairly safe that randomization will create equal groups. On the other hand, if you are using 5 or fewer participants in a group, randomization may not work. Part of the answer to the question of numbers boils down to what you feel comfortable with, what your research director feels comfortable with, or what you think you could defend to someone. In McClellan and Woods's (2001) study, there were 77 salesclerks divided into two groups. Given our guidelines, this number is quite adequate to use random assignment to create independent groups. On the other hand, although Wells (2001) had 41 students participate in the experiment, she may have decided to use repeated measures because she was worried about individual levels of stress that could have created a great deal of variability. Whatever you decide, it is critical to remember that the larger your samples, the more likely random assignment is to create equal groups.

Advantages of Correlated-Groups Designs There are two primary advantages correlated-groups designs provide to researchers: control and statistical issues. Both advantages are important to you as an experimenter.

Control Issues One basic assumption that we make before beginning our experiment is that the participants in our groups are equal with respect to the DV. When our groups are equal before the experiment begins and we observe differences between our groups on the DV after the experiment, then we can attribute those differences to the IV. Although randomization *should* equate our groups, the three methods for creating correlated-groups designs give us greater certainty of equality. We have exerted control to create equal groups. Thus, in correlated designs, we have some "proof" that our participants are equal beforehand. This equality helps us reduce some of the error variation in our experiment, which brings us to the statistical issues.

Statistical Issues Correlated-groups designs can actually benefit us statistically because they can help reduce error variation. You might be wondering, "What is error variation, anyhow?" In an experiment that involves one IV, you essentially have two sources of variability in your data. One source of variation is your IV: Scores on the DV should vary due to the two different treatment groups you have in your experiment. This source of variation, referred to as **between-groups variability**, is what you are attempting to measure in the experiment. Other factors that can cause variation in the DV, such as individual differences, measurement errors, and extraneous variation, are collectively known as **error variability**. As you might guess, our goal in an experiment is to *maximize* the between-groups variability and *minimize* the error or within-groups variability.

Between-groups variability Variability in DV scores that is due to the effects of the IV.

Error variability Variability in DV scores that is due to factors other than the IV, such as individual differences, measurement error, and extraneous variation (also known as *within-groups variability*).

Why is it important to reduce error variability? Although formulas for different statistical tests vary widely, they all reduce to the following general formula:

$$\text{statistic} = \frac{\text{between-groups variability}}{\text{error variability}}$$

Remember that the probability of a result occurring by chance goes down as the value of your statistic increases. Thus, larger statistical values are more likely to show significant differences in your experiment. Your knowledge of math tells you that there are two ways to increase the value of your statistic: *increase* the between-groups variability or

decrease the error variability. (Increasing between-groups variability is a function of your IV [see Chapter 6]; we will not discuss that option here.)

Can you figure out why using a correlated-groups design can reduce error variability?

Earlier in this section we listed individual differences as one source of error variability. Correlated-groups designs help reduce this source of error. If, in our treatment groups, we use the same participants or participants who share some important characteristic, either naturally or through matching, those participants will exhibit smaller individual differences between the groups than will randomly chosen participants. Imagine how dissimilar to you another participant could be if we chose that person at random. Imagine how similar to you another participant would be if that person were related to you, had the same intelligence as you, or (in the most obvious situation) were you! If we use a correlated design, then, error variability owing to individual differences should decrease, our statistic should increase, and we should have a greater chance of finding a significant difference as a result of our IV.

Why did we use the hedge word "should" three times in the preceding sentence?

Remember, when we discussed matched pairs earlier, we said that matching on an irrelevant variable could actually hurt your chances of finding a significant difference. If you match on an irrelevant variable, the between-groups differences do not decrease. If the between-groups differences do not decrease, your error variability is the same as if you had used an independent-groups design, which results in identical statistical test results. When we use a statistical test for a correlated-groups design, we must use a larger critical *t* value than we would have if we conducted the same experiment with an independent-groups design. (The statistical reason for this difference is that we give up some **degrees of freedom** in the correlated-groups design relative to an independent-groups design.) In the two-correlated-groups situation the degrees of freedom are equal to $N - 1$, where N represents the number of *pairs* of participants. In the two-independent-groups situation the degrees of freedom are $N - 2$, where N represents the *total* number of participants.

Degrees of freedom
The ability of a number in a specified set to assume any value.

Suppose you ran an experiment with 10 participants in each group. How many degrees of freedom would you have if this were an independent-groups design? A correlated-groups design?

Did you determine 18 *df* for the independent-groups design and 9 *df* for the correlated-groups design? Using the *t* table in the back of the book (see Table A-1), you will see that the critical *t* value at the .05 level with 18 *df* is 2.101, whereas it is 2.262 with 9 *df*.

These critical *t* values make it seem that it would be easier to reject the null hypothesis in the independent-samples situation (critical $t = 2.101$) than in the correlated-groups situation (critical $t = 2.262$). Yet we said earlier that the correlated-groups situation could benefit us statistically. What's going on?

The preceding numbers *do* support the first sentence of the previous paragraph—it *would* be easier to find a *t* of 2.101 than of 2.262. You must remember, however, that a correlated-groups design should reduce the error variability and result in a larger statistic. Typically, the statistic is increased more than enough to make up for the lost degrees of freedom. Remember that this reasoning is based on the assumption that you have matched on a relevant variable. Matching on an irrelevant variable does not reduce the error variability and will not increase the statistic—in which case, the lost degrees of freedom actually hurt your chances of finding significance. We will show you an actual statistical example of this point at the end of the statistical interpretation section later in this chapter.

Advantages of Independent-Groups Designs The chief advantage of independent-groups designs is their simplicity. Once you have planned your experiment, choosing your participants is quite easy—you merely get a large number of participants and randomly assign them to groups. You don't have to worry about measuring your participants on some variable and then matching them; you don't have to worry about whether each participant can serve in all conditions of your experiment; you don't have to worry about establishing or determining a relationship between your participants—these concerns are relevant only to correlated-groups designs.

Does the statistical advantage of correlated-groups designs render independent-groups designs useless? We cannot argue about the statistical advantage—it is real. However, as you can tell by reviewing the critical *t* values mentioned earlier, the advantage is not overwhelming. As the number of experimental participants increases, the difference becomes smaller and smaller. For example, the significant *t* value with 60 *df* is 2.00, and with 30 *df* it is only 2.04. If you expect your IV to have a powerful effect, then the statistical advantage of a correlated-groups design will be lessened.

One final point should be made in favor of independent-groups designs. Remember that, in some situations, it is simply impossible to use a correlated-groups design. Some circumstances do not allow repeated measures (as we pointed out earlier in the chapter), some participant variables cannot be matched, and some participants cannot be related in any way to other participants.

So what is the elusive conclusion? As you might guess, there is no simple, all-purpose answer. A correlated-groups design provides you with additional control and a greater chance of finding statistical significance. On the other hand, independent-groups designs are simpler to set up and conduct and can overcome the statistical advantages of correlated-groups designs if you use large samples. If you have large numbers of participants and expect your IV to have a large effect, you are quite safe with an independent-groups design. Alternatively, if you have only a small number of experimental participants and you expect your IV to have a small effect, the advantages of a correlated-groups design would be important to you. For all those in-between cases you must weigh the alternatives and choose the type of design that seems to have the greater advantage.

Variations on the Two-Group Design

To this point we have described the two-group design as if all two-group designs were identical. This is not the case. Let's look at two variations on this design.

Comparing Different Amounts of an IV Earlier we said that the most common use of two-group designs was to compare a group of participants receiving the IV (experimental group) to a group that does not receive the IV (control group). Although this is the most common type of two-group design, it is not the only type. The presence–absence manipulation of an IV allows you to determine whether the IV has an effect. For example, McClellan and Woods (2001) were able to determine that customers with a hearing disability received help from salesclerks more slowly than customers without such a disability; Wells (2001) found that her students did react to a mental arithmetic task in a stressful manner. An analogous situation to using the presence–absence manipulation in detective work is trying to sort out the clues that relate to the guilt or innocence of a single suspect.

A presence–absence IV manipulation does not, however, allow you to determine the precise effects of the IV. McClellan and Woods did *not* discover whether having salesclerks wait on deaf customers caused a large or small delay in offering help, only that the clerks responded more slowly. Wells did not determine how stressful mental arithmetic was compared to other tasks, only that it did increase stress.

Typically, after we determine that a particular IV has an effect, we would like to have more specific information about that effect. Can we produce the effect with more (or less) of the IV? Will a different IV produce a stronger (or weaker) effect? What is the optimum amount (or type) of the IV? These are just a few of the possible questions that remain after determining that the IV had an effect. Thus, we can follow up on our IV presence–absence experiment with a new two-group experiment that compares different *amounts* or *types* of the IV to determine their effectiveness. Similarly, a detective may have two suspects and be faced with the task of sorting out the evidence to decide which suspect is more likely the guilty party.

Some IVs simply cannot be contrasted through presence–absence manipulations. For example, Erin Vaughn (2002) of Ouachita Baptist University in Arkadelphia, Arkansas, was interested in the effects of stereotyping on "person perception." In her experiment Vaughn wanted to assess the effects of the tendency to stereotype people on the basis of physical attractiveness (IV) on pairing pictures of men and women into dating couples (DV). This question would make little sense in an IV presence–absence situation: People could show more or less of a tendency to stereotype, but not zero or total tendency. Vaughn therefore compared differing amounts of this IV. Based on participants' answers to a survey, she formed "high tendency to stereotype" and "low tendency to stereotype" groups. When she compared these two groups' pairings of men and women, she found that people with a high tendency to stereotype created more pairings in which men and women were similar in physical attractiveness. Thus, the tendency to stereotype affected the way that participants believed that dating couples "paired up" in Vaughn's experiment.

The key point to notice when we conduct an experiment contrasting different amounts of our IV is that we no longer have a true control group. In other words, there is no group that receives a zero amount of the IV. Again, we are not trying to determine whether the IV has an effect—we already know that it does. We are merely attempting to find a difference between differing types or amounts of our IV.

Dealing with Measured IVs To this point, when we have mentioned IVs in this text, we have emphasized that they are the factors that the experimenter *directly manipulates*.

Technically, this statement is correct only for a **true experiment**. In a true experiment the experimenter has total control over the IV and can assign participants to IV conditions. In other words, the experimenter can manipulate the IV. McClellan and Woods (2001) were able to assign their clerks to either the deaf-customer group or the hearing-customer group. If you wish to assess the effects of two different reading programs on teaching children how to read, you can assign nonreading children to either program.

True experiment An experiment in which the experimenter directly manipulates the IV.

As you saw in Chapter 6, there are many IVs (participant IVs) that psychologists wish to study but cannot directly manipulate; we measure them instead. For example, Lindsey Smith, a student at Presbyterian College in Clinton, South Carolina, and Marion Gaines, her faculty sponsor, examined performance differences on a perceptual task between college students with attention deficit hyperactivity disorder (ADHD) and students without ADHD. Because they were not able to directly manipulate the ADHD status of their participants (of course), Smith and Gaines (2005) conducted **ex post facto research** (see Chapter 4). In the context of the present chapter they used a two-group design. The participants completed a backward masking perceptual task in which they viewed stimuli on a computer screen that could be masked with Xs to make the perception more difficult. Smith and Gaines found that the participants with ADHD performed more poorly on the perceptual task than the control group. Based on what we know about ADHD, Smith and Gaines's finding may not be surprising. We can even develop hypotheses about why the students with ADHD tended to perform more poorly on the perceptual task; however, we must be cautious in our interpretation. Although we do know from Smith and Gaines's research that participants without ADHD performed better than participants with ADHD, we are not certain why this difference exists. In other words, there could be other differences in the two groups of participants than simply the ADHD. Because Smith and Gaines did not (*could* not) assign participants to groups randomly (the definition of ex post facto research), they could not be certain that ADHD was the only difference between the groups.

Ex post facto research A research approach in which the experimenter cannot directly manipulate the IV but can only classify, categorize, or measure the IV because it is predetermined in the participants (e.g., IV = sex).

The drawback of ex post facto research is certainly a serious one. Conducting an experiment without being able to draw a cause-and-effect conclusion is limiting. Why would we want to conduct ex post facto research if we cannot draw definitive conclusions from it? As we mentioned earlier, some of the most interesting psychological variables do not lend themselves to any type of research other than ex post facto. If you wish to study the genesis of female–male differences, you have no option other than conducting ex post facto studies. Also, as psychologists continue to conduct ex post facto research, they do make progress. Attempting to specify the determinants of intelligence involves ex post facto research—surely you remember the famous heredity-versus-environment debate over IQ. What we *think* we know today is that both factors affect IQ: Psychologists believe that heredity sets the limits of your possible IQ (i.e., your possible minimum and maximum IQs) and that your environment determines where you fall within that range (Weinberg, 1989). Thus, it seems clear that we should not abandon ex post facto research despite its major drawback. We must, however, remember to be extremely cautious in drawing conclusions from ex post facto studies. Detectives, of course, are always faced with ex post facto evidence—it is impossible to manipulate the variables *after* a crime has been committed.

REVIEW SUMMARY

1. **Correlated-groups designs** provide more control because they guarantee equality of the two groups.
2. Correlated-groups designs generally reduce **error variability** and are more likely to achieve statistically significant results.
3. An advantage of **independent-groups designs** is that they are simple to conduct. With large numbers of research participants, they are also strong designs.
4. Researchers often use two-group designs to compare different amounts (or types) of IVs.
5. We cannot manipulate some IVs, so we must resort to measuring them and conducting **ex post facto research**, which cannot demonstrate cause-and-effect relations.

Check Your Progress

1. Why is it important that our two groups be equal before the experiment begins?
2. The variability in DV scores that can be attributed to our experimental treatments is called _____ _____; variability from other sources is labeled _____ _____.
3. Which three factors cause variation in DV scores (error variability)?
   a. nonrandom assignment, random assignment, and mixed assignment
   b. individual differences, measurement errors, and extraneous variables
   c. placebo effects, measurement errors, and the IV
   d. variance, standard deviation, and the mean
4. Compare and contrast the advantages and disadvantages of independent-groups and correlated-groups designs.
5. Other than the consent form and mental arithmetic IV examples given in the chapter, give two examples of IVs for which you might wish to compare differing amounts.
6. List three examples of IVs not in the text that you would have to study with ex post facto experiments.

Statistical Analysis: What Do Your Data Show?

After you have used your experimental design to conduct an experiment and gather data, you are ready to use your statistical tools to analyze the data. Let's pause for a moment to understand how your experimental design and statistical tests are integrated.

The Relation Between Experimental Design and Statistics

At the beginning of this chapter we compared experimental design to a blueprint and pointed out that you needed a design to know where you were headed. When you carefully plan your experiment and choose the correct experimental design, you also accomplish

another big step. Selecting the appropriate experimental design determines the particular statistical test you will use to analyze your data. Because experimental design and statistics are intimately linked, you should determine your experimental design *before* you begin collecting data to ensure there will be an appropriate statistical test you can use to analyze your data. Remember, you don't want to be a professor's classroom example of a student who conducted research project only to find out there was no way to analyze the data!

Analyzing Two-Group Designs

In this chapter we have looked at one-IV, two-group designs. You may remember from your statistics course, as well as from Chapter 9, that this type of experimental design requires a *t* test to analyze the resulting data (assuming you have interval- or ratio-level data). You may also remember learning about two different types of *t* tests in your statistics class. For a two-independent-groups design you would use a *t* test for independent samples (also known as an independent *t* test) to analyze your data. For a two-correlated-groups design you would analyze your data with a *t* test for correlated samples (also called a dependent *t* test, a within-groups *t* test, or a paired *t* test).

Let's make certain that the relation between experimental design and statistics is clear. A *t* test is indicated as the appropriate statistical test because you conducted an experiment with one IV that has two levels (treatment conditions). The decision of *which t* test to use is based on how you assigned your participants to their groups. If you used random assignment, then you will use the *t* test for independent samples. If you used repeated measures, matched pairs, or natural pairs, then you would use the *t* test for correlated samples.

Calculating Your Statistics

In Chapter 9 we provided the computer analysis of a *t* test. The research example involved a comparison of how long it took salespeople to wait on customers who were dressed in sloppy or dressy clothes. In this chapter we will examine those data more completely. To help set the stage for the remainder of the chapter, let's review some details of the hypothetical experiment behind the data. We wondered whether the clothes students wore actually make any difference in how quickly salesclerks would wait on them. We collected data from 16 different clerks, randomly assigning 8 to wait on customers wearing dressy clothes and 8 to wait on customers wearing sloppy clothes.

Which statistical test would you use to analyze data from this experiment and why?

The simplest way to answer these questions is to use our chart in Figure 10-1. There is only one IV: the type of clothing worn. That IV has two levels: dressy and sloppy. We randomly assigned the participants to their groups. Thus, this design represents a two-independent-groups design, and you should analyze it with an independent *t* test.

Interpretation: Making Sense of Your Statistics

We hope that your statistics instructor taught you this important lesson about statistics: Statistics are not something to fear and avoid; they are a tool to help you understand the data garnered from your experiment. Because of today's focus on computerized statistical analyses, calculating statistics is becoming secondary to interpreting them. Just as having a sewing machine is useless if you don't know how to operate it, statistics are useless if you don't know how to interpret them. Likewise, detectives must learn the skills necessary to interpret the reports they receive from the police scientific labs. We will focus on two types of interpretation in this section: interpreting computer statistical output and translating statistical information into experimental outcomes.

Interpreting Computer Statistical Output

There may be hundreds of computer packages available for analyzing data. Thus, it would be impossible (and inefficient) to show output from every different package and teach you how to interpret each one. Remember that we will show you generic computer statistical output and present interpretations of those analyses. We believe that the similarity among statistical packages will allow you to generalize from our examples to the specific package that you may use. (Computerized statistical packages vary widely in the number of decimal places they report for statistical results. To be consistent with APA format, we will round the computerized output and use only two decimal places in the text.)

The *t* Test for Independent Samples Let's return to our statistical example from Chapter 9. Remember, we randomly assigned clerks to one of two groups: a sloppily dressed group of customers or a well-dressed group. We sent the customers to stores and obtained the time-to-service scores you saw in Chapter 9. If we analyzed these data using a computer package, what might the output look like? We presented an abbreviated version of the output in Chapter 9 for simplicity's sake; a more complete printout appears in Table 10-1.

TABLE 10-1 Computer Output for *t* Test for Independent Groups

GROUP 1 = Dressy clothing
GROUP 2 = Sloppy clothing
Variable = Salesclerks' response time

GROUP	*N*	Mean	*SD*	Standard Error
GROUP 1	8	48.38	10.113	3.575
GROUP 2	8	63.25	12.544	4.435

F_{max}test *F* = 1.634 *p* = 0.222

Equal Variances Assumed

t = 2.61 *df* = 14 *p* = 0.021 Cohen's *d* = 0.92

Equal Variances Not Assumed

t = 2.61 *df* = 13.4 *p* = 0.021 Cohen's *d* = 0.92

We usually examine the descriptive statistics first. The descriptive statistics are printed at the top of the printout. We see that GROUP 1 (defined at the top of the printout as "Dressy Clothing") had 8 cases, a mean salesperson response time of 48.38 seconds, a standard deviation of 10.11 seconds, and a standard error of 3.58 seconds. GROUP 2 (the "Sloppy Clothing" group) had 8 cases, a mean salesperson response time of 63.25 seconds, a standard deviation of 12.54 seconds, and a standard error of 4.44 seconds. Be cautious at this point—an old saying we learned regarding computers is "garbage in, garbage out." In other words, if you enter incorrect numbers into a computer, you will get incorrect numbers out of the computer. You should always verify any numbers you enter and, as much as possible, double-check the output. "Wait a minute," you may be saying. "What's the use of using a computer if I have to check its work?" We're not suggesting that you check up on the computer but that you check up on yourself! For example, suppose the computer information for GROUP 1 or GROUP 2 showed the number of cases to be seven. You would know that the computer didn't read one number—perhaps you entered only seven scores, or perhaps you mislabeled one score. With only eight scores, it is simple enough to calculate the mean for each group yourself. Why should you do that? If you find the same mean that the computer displays, you can be reasonably certain that you entered the data correctly and, therefore, can go on to interpret your statistics.

The second set of statistics provided contains only two statistical values: F and p. These values represent the results of a test known as F_{max}, a statistic used to test the assumption of **homogeneity of variance** for the two groups (Kirk, 1968). Homogeneity of variance simply means that the variability of the scores of the two groups is similar. To use a t test, we must assume that the variances are similar. In this particular example our assumption is justified because the probability of chance for the F value is .22, well above the standard .05 cutoff. Because we have homogeneity of variance, we will use the third block of information ("Equal Variances Assumed") to interpret our test.

Homogeneity of variance The assumption that the variances are equal for the two (or more) groups you plan to compare statistically.

In the second set of information, if our p value were less than .05, we would have found **heterogeneity of variance**, meaning that the variability of the scores of the two groups was *not* comparable. Thus, we would be violating a mathematical assumption for using the t test. Fortunately, statisticians have developed a procedure that allows us to interpret our statistics despite heterogeneity. In such a case we would use the fourth block of information ("Equal Variances Not Assumed") rather than the third block. If the variances of the two groups are equivalent, we can pool or combine those estimates; however, if the variances are not equivalent, we must keep them separate. Again, in our current example, because the F_{max} statistic is not significant ($p = .22$), we will use the statistical results under the "Equal Variances Assumed" heading.

Heterogeneity of variance Occurs when we do not have homogeneity of variance; this means that our two (or more) groups' variances are not equivalent.

Generally speaking, t tests are **robust** with regard to the assumption of homogeneity (Kirk, 1968). A robust test is one that can tolerate violations of its assumptions and still provide accurate answers. Kirk noted that the t test is so robust that the homogeneity assumption is often not even tested. The statistics package you use, therefore, may not provide information about the F_{max} statistic (and thus probably will not give you equal and unequal variance estimates).

Robust Refers to a statistical test that can tolerate violation of its assumptions (e.g., homogeneity of variances) and still yield valid results.

By looking at the third set of information, we find that our *t* value (calculated by the computer) is 2.61. We have 14 degrees of freedom ($N_1 + N_2 - 2$). Rather than having to locate these values in a *t* table to determine significance, we can use the significance level provided by the computer: The probability (two-tail) is .021. Thus, the probability that two means as different as these could have come from the same population by chance is less than 3 in 100. This probability is less than the magical .05 cutoff, so we conclude that these two means are significantly different (i.e., the difference between them is *not* due to chance).

Some statistical packages may not automatically print the degrees of freedom for you, so it is important to remember how to calculate *df.* Also, some programs may not provide the probability of your result as part of the printout; then you would have to make this determination yourself. In such a case you would use the *t* table (Appendix A, Table A-3). In this case you would find that the probability of the *t* we found is less than .05. (Computer output typically provides exact *p* values [.021 in this case], whereas statistical tables simply allow you to compare your result to standard *p* values such as .05 or .01.

In addition, the computer output shows that Cohen's *d* is 0.92. We learned in Chapter 9 that a *d* of 0.8 or larger is considered a large effect size. This information helps to confirm that customers' attire plays a major role in determining salesclerks' speed of helping. This decision completes the process of interpreting the computer output. Our next task is to describe our statistical information in terms of the experiment we conducted.

Translating Statistics Into Words Think back to the logic of an experiment: We start an experiment with two equal groups and treat them identically (for control purposes) with one exception (our IV, or type of dress); we measure the two groups (on our DV, or time to provide service) in order to compare them. At this point, based on our statistical analyses, we know that we have a significant difference (i.e., not due to chance) between our two means. If two equal groups began the experiment and they are now *un*equal, to what can we attribute that difference? If our controls have been adequate, our only choice is to assume that the difference between the groups is due to the IV.

Looking at our example, we have decided that the groups of students dressed in two different types of clothing received help from clerks in different amounts of time. Many students stop at this point, thinking that they have drawn a complete conclusion from their experiment.

Why would this conclusion be incomplete? Can you develop a complete conclusion before proceeding?

Saying that students who are dressed differently get waited on in different amounts of time is an incomplete conclusion because it specifies only a difference, not the *direction* of that difference. Whenever we compare treatments and find a difference, we want to know which group has performed at a better or higher level. In a two-group experiment this interpretation is quite simple. Because we have only two groups and we have concluded that they differed significantly, we can further conclude that the group with the higher mean score has outscored the group with the lower mean score (remember that high scores do not always indicate superior performance, as in this case).

To interpret fully the results from this experiment, we examine our descriptive statistics and find that the salespeople waiting on students dressed in sloppy clothes had a mean response time of 63.25 seconds, whereas the salespeople waiting on the well-dressed students averaged 48.38 seconds. Thus, we can conclude that the salesclerks waiting on nicely dressed customers responded more quickly than those waiting on sloppily dressed customers. Notice that this statement includes both the notion of a difference *and* the direction of that difference.

When we draw conclusions from our research, we want to communicate those results clearly and concisely in our experimental report. To accomplish these two objectives, we use both words and numbers in our communication. This communication pattern is part of the APA style for preparing research reports, which we will consider in Chapter 14 (APA, 2001). We will introduce the form for statistical results here. Bear in mind that you are trying to communicate—to tell what you found in words and provide statistical information to support those words. For example, if you were writing an interpretation of the results from our sample experiment, you might write something like the following:

> Salesclerks who waited on well-dressed customers ($M = 48.38, SD = 10.11$) took significantly less time, $t(14) = 2.61, p = .021$, to respond to customers than salespeople who waited on customers dressed in sloppy clothing ($M = 63.25$, $SD = 12.54$). The effect size, estimated with Cohen's d, was .92.

Notice that the words alone give a clear account of the findings—a person who has never taken a statistics course could understand this conclusion. The inferential statistics regarding the test findings support the conclusion. The descriptive statistics (M = mean, SD = standard deviation) given for each group allow the reader to see how the groups actually performed and to see how variable the data were. This standard format allows us to communicate our statistical results clearly and concisely.

The *t* Test for Correlated Samples Remember that we have covered two different two-group designs in this chapter. Now we will examine the computer output for analysis of the two-correlated-groups design. Our experiment concerning the salespeople was an example of the two-independent-groups design, which would *not* require a *t* test for correlated samples.

How could you modify this experiment so that it used correlated groups rather than independent groups?

You should remember that there are three methods for creating correlated groups: matched pairs, repeated measures, and natural pairs. If your modified experiment used one of these techniques, you made a correct change. As an example, let's assume that we were worried about the difference between salesclerks confounding our experiment. To better equate the clerks in our two groups, we decide to use the repeated-measures approach. We decide to measure each salesclerk's time to respond to two customers: once for a dressed-up customer and once for a sloppily dressed customer. Before beginning our experiment, we know that the salespeople are identical for the two groups, thus removing individual differences as a potential confounding variable.

Next, we conduct our experiment. We measure the response time of each of the eight clerks waiting on both types of customers (based on dress). Given this hypothetical example, the

TABLE 10-2 Computer Output for *t* Test for Correlated Groups

```
GROUP 1 = Dressy clothing
GROUP 2 = Sloppy clothing
Variable = Salesclerks' response time

                N        Mean          SD          Standard Error
GROUP 1         8       48.38       10.113              3.575
GROUP 2         8       63.25       12.544              4.435

Mean difference = 14.875    SD = 7.699      Std Error = 2.722

Corr. = 0.790       p = 0.020

t = 5.465    df = 7         p = 0.001      Cohen's d = 1.93
```

scores from Chapter 9 would now represent repeated-measures time scores rather than independent scores. After analyzing the data with our computer package, we find the output in Table 10-2. (Please note that it is *not* legitimate to analyze the same data with two different statistical tests. We are doing so in this chapter merely for example's sake. If you tested real-world data multiple times, you would increase the probability of making a Type I error; see Chapter 9.)

Look at Table 10-2. Again, we first look for the descriptive statistics and find them at the top of the printout. Of course, because we used the same data, we have the same descriptive statistics as for the independent-samples test. Salesclerks waiting on the students wearing sloppy clothing responded in an average of 63.25 seconds, with a standard deviation of 12.54 and a standard error of 4.44. The students who wore dressy clothes received help in 48.38 seconds, with a standard deviation of 10.11 and a standard error of 3.58. Remember that there are 8 *pairs* of scores (representing the 8 clerks) rather than 16 individual scores. This difference between the two *t* tests will be important when we consider the degrees of freedom.

The second block of information shows us the size of the difference between the two means, as well as its standard deviation and standard error. (Researchers rarely use this information, so it may not appear in your computer output.) The third block gives you some information about the relation between the pairs of participants (or the same participant for repeated measures). Here you can determine whether the paired scores were correlated. Remember that we want them to be correlated so that we will gain the additional statistical control made available by using the correlated-groups design. As you can see in this example, the scores were highly **positively correlated** (see Chapters 4 and 9). In our example, this result implies that if a salesclerk waited on one student quickly, he or she tended also to wait on the other student quickly.

Positive correlation As scores on one variable increase, scores on the second variable also increase.

In the fourth block we find the results of our inferential test. We obtained a *t* value of 5.47 with 7 degrees of freedom.

We have 16 data points (8 clerks measured twice each) in our experiment but only 7 degrees of freedom. In our earlier example, we had 16 participants and 14 degrees of freedom. What is the difference in this case?

You should remember that our degrees of freedom for correlated-samples cases are equal to the *number of pairs of participants minus 1*. If this is fuzzy in your memory, refer to Statistical Issues earlier in the chapter.

The computer tells us that the probability of a t of 5.47 with 7 df is .001. With such a low probability of chance for our results, we would conclude that there is a significant difference between the clerks' response times to differently dressed students. In other words, we believe that it is highly unlikely that the difference between our groups could have occurred by chance and that, instead, the difference must be due to our IV. The effect size information, Cohen's d, provides ample support for our conclusion as d is 1.93. Remember that 0.8 represents a large effect size; therefore, the effect of the IV is quite substantial in this analysis.

Translating Statistics Into Words Our experimental logic is exactly the same for this experiment as it was for the independent-samples case. The only difference is that with our matched participants, we are more certain that the two groups are equal before the experiment begins. We still treat our groups equally (control) with the one exception (our IV) and measure their performance (our DV) so that we can compare them statistically.

To translate our statistics into words, it is important to say more than the fact that we found a significant difference. We must know what form or direction that significant difference takes. With the t test for correlated samples, we are again comparing two groups, so it is a simple matter of looking at the group means to determine which group outperformed the other. Of course, because we are using the same data, our results are identical: The sloppily dressed students received help in a mean of 63.25 seconds compared to 48.38 seconds for the well-dressed students.

How would you write the results of this experiment in words and numbers for your experimental report?

Did you find yourself flipping back in the book to look at our earlier conclusion? If so, that's a good strategy because this conclusion should be quite similar to the earlier conclusion. In fact, you could almost copy the earlier conclusion as long as you made several important changes. Did you catch those changes? Here's an adaptation of our earlier conclusion:

> Salespeople who waited on well-dressed customers (M = 48.38, SD = 10.11) took significantly less time, $t(7)$ = 5.47, p = .001, to respond to the customers than when they waited on customers dressed in sloppy clothes (M = 63.25, SD = 12.54). The effect size, estimated with Cohen's d, was 1.93.

As you can see, four numbers in the sentences changed: We had fewer degrees of freedom, our t value was larger, our probability of chance was lower, and our effect size was much larger. In this *purely hypothetical* example in which we analyzed the same data twice (a clear violation of assumptions if you were to do this in the real world), you can see the advantage of correlated-groups designs. Although we lost degrees of freedom compared to the independent-samples case presented earlier in the chapter, the probability that our results were due to chance

actually decreased and our effect size increased dramatically. Again, we gained these advantages because our matching of the participants decreased the variability in the data.

You can see a vivid illustration of what we gained through matching by examining Table 10-3. In this example we have used the same data from Chapter 9 but shuffled the scores for the second group before we ran a *t* test for correlated samples. Such a situation would occur if you matched your participants on an irrelevant variable (remember that we mentioned this possibility earlier in the chapter). As you can see by comparing Tables 10-3 and 10-2, the descriptive statistics remained the same because the scores in each group did not change. However, in Table 10-3 the correlation between the two sets of scores is now −.88. Because there is a **negative correlation** between our scores, the *t* value is 1.91, even lower than it was in our *t* test for independent groups (see Table 10-1). The marked change comes when we compare the inferential statistics in Tables 10-2 and 10-3. The original analysis showed a *t* of 5.47 with $p = .001$. In contrast, with a negative correlation between the pairs of scores, the new analysis shows a *t* of 1.91 with $p = .097$ and an effect size of 0.68 (the smallest of our three analyses). Thus, these results did not remain significant when the correlation between the participants disappeared. Again, the key point to remember is that when using a correlated-groups design, the groups should actually be positively correlated.

Negative correlation
As scores on one variable increase, scores on the second variable decrease.

The Continuing Research Problem

Research is a cyclical, ongoing process. It would be rare for a psychologist to conduct a single research project and stop at that point because that one project had answered all the questions about the particular topic. Instead, one experiment usually answers some of your questions, does not answer others, and raises new ones for your consideration. As you have studied the work of famous psychologists, you may have noticed that many of them established a research area early in their careers and continued working in that area for the duration of their professional lives. We're not trying to say that the research area you choose as an undergraduate will shape your future as a psychologist—although it could! Rather, we are merely pointing out that one good experiment often leads to another.

TABLE 10-3 Computer Output for *t* Test for Correlated Groups (Shuffled Data)

```
GROUP 1 = Dressy clothing
GROUP 2 = Sloppy clothing
Variable = Salesclerks' response time

Group          N        Mean          SD          Standard Error

GROUP 1        8        48.38       10.113              3.575
GROUP 2        8        63.25       12.544              4.435

Mean difference = 14.88     SD = 21.977     Std Error = 7.770

Corr. = -.880       p = 0.004

t = 1.91     df = 7     p = 0.097     Cohen's d = 0.68
```

We want to show you how research is an ongoing process as we move through the next two chapters with our continuing research problem. We sketched out a research problem in Chapter 9 (comparing how customers' style of dress affects salespeople's performance) and asked you to help us solve the problem through experimentation. We will continue to examine this problem throughout the next two chapters so that you can see how different questions we ask about the same problem may require different research designs. This research problem is purely hypothetical, but it has an applied slant to it. We hope the continuing research problem helps you see how a single question can be asked in many different ways *and* that a single question often leads to many new questions.

To make certain you understood the logical series of steps we took in choosing a design, let's review those steps, paying particular attention to the experimental design questions shown in Figure 10-1:

1. After reviewing relevant research literature, we chose our IV (style of dress) and our DV (salesclerk response time).
2. Because we were conducting a preliminary investigation into the effects of clothing on salesclerks' reactions, we decided to test only one IV (the style of dress).
3. Because we wanted to determine only whether clothing style can affect the performance of salespeople, we chose to use only two levels of the IV (dressy clothing vs. sloppy clothing).

4a. If we have a large number of participants available, then we can use random assignment, which yields independent groups. In this case we would use the two-independent-groups design and analyze the data with a *t* test for independent groups.

4b. If we expect to have a small number of participants and must exert the maximum degree of control, we choose to use a design with repeated measures or matched groups, thus resulting in correlated groups. Therefore, we would use a two-correlated-groups design for the experiment and analyze the data with a *t* test for correlated groups.

5. We concluded that salespeople responded more quickly to customers in dressy clothes than to customers dressed in sloppy clothes.

REVIEW SUMMARY

1. The statistical test you use for analyzing your experimental data is related to the experimental design you choose.
2. When you have one IV with two groups and use randomly assigned research participants, the appropriate statistical test is the ***t* test for *independent* samples**.
3. When you have one IV with two groups and use matched pairs, natural pairs, or repeated measures with your participants, the appropriate statistical test is the ***t* test for *correlated* samples**.
4. Computer printouts of statistics typically give descriptive statistics (including means and standard deviations) and inferential statistics.
5. To communicate the statistical results of an experiment, we use APA format for clarity and conciseness.
6. Research is a cyclical, ongoing process. Most experimental questions can be tested with different designs.

Check Your Progress

1. Which statistical test would you use if you compared the stereotyping of a group of female executives to a group of male executives? Explain your reasoning.
2. Which statistical test would you use if you compared the stereotyping of a group of male executives before and after an antidiscrimination bill passed through Congress? Explain your reasoning.
3. Compared to the *t* test for independent groups, the *t* test for correlated samples has _____ degrees of freedom.
   a. fewer
   b. more
   c. exactly the same number of
   d. none of these; it is impossible to compare degrees of freedom between statistical tests
4. What information do we usually look for first on a computer printout? Why?
5. If the variability of our two groups is similar, we have _____; if the variability of the groups is dissimilar, we have _____.
6. When we write a report of our experimental results, we explain the results in _____ and _____.
7. Interpret the following statistics:

   Group A ($M = 75$); Group B ($M = 70$); $t(14) = 2.53$, $p < 0.05$
8. Why do we describe research as a cyclical, ongoing process? Give an example of how this cycling might take place.

Key Terms

Experimental design, 203
Principle of parsimony, 204
Independent variable (IV), 204
Dependent variable (DV), 204
Extraneous variables, 206
Levels, 206
Experimental group, 206
Control group, 206
Random assignment, 207
Random selection, 208
Independent groups, 208
Between-subjects comparison, 208
Confounded experiment, 208
Correlated assignment, 209
Matched pairs, 210
Repeated measures, 211
Natural pairs, 212
Within-subjects comparison, 212
Between-groups variability, 215
Error variability, 215
Degrees of freedom, 216
True experiment, 219
Ex post facto research, 219
Homogeneity of variance, 223
Heterogeneity of variance, 223
Robust, 223
Positive correlation, 226
Negative correlation, 228

Looking Ahead

In this chapter we have examined the notion of planning an experiment by selecting a research design. In particular we examined the basic building-block designs with one IV and two groups. In the next chapter we will enlarge this basic design by adding more groups to our one IV. This enlarged design will give us the capability to ask more penetrating questions about the effects of our IV and to obtain more specific information about those effects.