CHAPTER 10

The Simple Experiment

What you have is an experience, not an experiment.
—R. A. Fisher

Happy is the person who gets to know the reasons for things.
—Virgil

CHAPTER OVERVIEW

Why do people behave the way they do? How can we help people change? To answer these questions, we must be able to isolate the underlying causes of behavior, and to do that, we must design a study that has **internal validity**: the ability to determine whether a factor causes an effect.

This chapter introduces you to one of the easiest ways to establish that a factor causes an effect: the simple experiment. You will start by learning the basic logic behind the simple experiment. Then, you will learn how to weigh statistical, ethical, and validity issues in order to design a useful simple experiment. Finally, you will learn how to interpret the results of such an experiment.

LOGIC AND TERMINOLOGY

The **simple experiment** involves two groups of participants. At the start of the experiment, the two groups should not differ from each other in any systematic way, but during the experiment, the experimenter will treat one group differently from the other. For example, the experimenter may

- Assign the groups different *types* of activities (e.g., playing violent versus nonviolent video games)
- Assign the groups different *amounts* of an activity (e.g., one group might meditate for 30 minutes whereas the other group meditates for 10 minutes)
- Appear one way (e.g., well dressed) to one group and another way (e.g., casually dressed) to the other group
- Have confederates (people who pretend to be participants but who are actually the researcher's assistants) behave one way (e.g., agreeing with the participant) when interacting with members of one group and another way (e.g., disagreeing with the participant) when interacting with the other group
- Have a certain object (e.g., a mirror or a gun) in the testing room when members of one group are tested but not when members of the other group are tested
- Make the testing room's environment more intense on a certain dimension (e.g., how hot it is, how loud it is, how it is lit, how it smells, or the concentration of negative ions in it) when members of one group are tested and less intense on that dimension when the other group is tested

- Give the groups different instructions ("memorize these words by repeating them over and over" versus "make a sentence out of these words," or "keep a log of what you have to be grateful for" versus "keep a log of hassles you encounter")
- Give the groups different printed stimuli (whether or not the sentences participants are asked to unscramble make participants think about older people, whether the words participants are to memorize are concrete and easy to visualize [e.g., "bell"] or abstract and hard to visualize [e.g., "liberty"], whether the exam is printed on blue or white paper, whether the people in the photograph are attractive or unattractive)
- Give the groups different contexts for interpreting stimuli (the researcher may vary the gender, age, attractiveness, or background of the person whose job application, school record, essay, or character is being judged)
- Give the groups different scenarios (the situations may be the same but worded differently [e.g., "Valerie and I are best friends" versus "We are best friends" or "You can have $5.00 now or $6.20 in a month" versus "You can have $5.00 now and $0 in a month or $0 now and $6.20 in a month"] or the scenarios may differ in one respect (e.g., gender, race, or job experience of characters; the possible or likely causes of an event [e.g., the person was—or was not—drunk, the disease could—or could not—be transmitted through sexual contact])
- Give the groups different feedback ("the test suggests you are outgoing" versus "the test suggests you are shy," "the test suggests you will spend much of your future alone" versus, "the test suggests you will spend much of your future with friends and loved ones" or "you did well on the task" versus "your performance on the task was average)
- Give the groups different chemicals (sugar-sweetened lemonade versus artificially sweetened lemonade, caffeinated versus decaffeinated colas)

Often, half the participants (the treatment group) receive a treatment, whereas the other half (the no-treatment group) receive no treatment. If, at the end of the experiment, the two groups differ significantly, we can conclude that the treatment—the only systematic difference between the groups—caused that significant difference.

But how do we set up a situation in which the only systematic difference between the no-treatment and the treatment groups is the treatment? The answer is **independent random assignment**. In random assignment, a process similar to determining what treatment the participant will receive based on a coin flip, every participant—regardless of that participant's characteristics—has an equal chance of being assigned to either the treatment or no-treatment group. If we provide each participant an equal chance of being assigned to either group, there will still be unsystematic, chance differences between our groups before we introduce the treatment, but there should *not* be any *systematic* differences between them.

To review, random assignment, the key to the simple experiment, involves two processes. First, we *randomly* divide our participants into two similar halves. Second, we *assign* one of those halves to get a treatment different from the other. For example, half may be allowed to choose the deadlines for their term papers, whereas the other half are not; or half the participants would be given a violent video game to play, whereas the other half would be given a neutral video game.

We have given you a general idea of what random assignment is, but how would you actually randomly assign participants to either a no-treatment or a treatment group?[1] You might think that you could flip a coin for each participant: If the coin comes up heads, the participant gets the treatment; if the coin comes up tails, the participant does not get the treatment. However, coin-flipping does not work because "a tossed coin is slightly more likely to land on the face it started out on than on the opposite face" (Klarreich, 2004, p. 363). Even computers have trouble producing random sequences (Klarreich, 2004). So what should you do? (The "eenie meenie minie moe" method is not an option because it isn't random.) The solution is to use a random numbers table to assign participants to condition (Wilkinson & the Task Force on Statistical Inference, 1999). To learn how to use a random numbers table, see Box 10.1.

Experimental Hypothesis: The Treatment Has an Effect

If you do not randomly assign your participants to two groups, you do not have a simple experiment. However, before you randomly assign participants, you must have an **experimental hypothesis**: a prediction that the treatment will cause an effect. To generate an experimental hypothesis, you must predict that the treatment and no-treatment groups will differ because of the treatment's effect. For example, you might hypothesize that participants getting 3 hours of full-spectrum light will be happier than those getting no full-spectrum light because full-spectrum light causes increases in happiness.

Although you can make a wide variety of experimental hypotheses (e.g., you could hypothesize that participants forced to trade their lottery tickets would be unhappier than those who were not forced to trade their lottery tickets or that participants forced to describe their relationship with their friend with "My friend and I _____" sentences would be less happy with the relationship than people forced to describe their relationship with "We_____" sentences), realize that not all hypotheses are cause–effect hypotheses. Sometimes, hypotheses involve describing what happens rather than finding out what makes things happen. If you generate a hypothesis that is *not* a cause–effect statement, it is *not* an experimental hypothesis. Thus, if you hypothesize that men are more romantic than women, you do not have an experimental hypothesis. Similarly, if you predict that athletes will be more assertive than nonathletes, you do not have an experimental hypothesis. In short, to have an experimental hypothesis, you must predict that some *treatment* that you manipulate will *cause* an *effect*.

Null Hypothesis: The Treatment Does Not Have an Effect

Once you have an experimental (cause–effect) hypothesis, pit it against the **null hypothesis**: the hypothesis that the treatment has *no* effect. The null hypothesis essentially states that any difference you observe between the treatment and no-treatment group scores could be due to chance. Therefore, if our experimental hypothesis was that getting 3 hours of full-spectrum lighting

[1] Instead of using pure independent random assignment, researchers typically use independent random assignment with the restriction that an equal number of participants must be in each group.

BOX 10.1 Randomly Assigning Participants to Two Groups

There are many ways to randomly assign participants to groups. Your professor may prefer another method. However, following these steps guarantees random assignment and an equal number of participants in each group.

Step 1: On the top of a sheet of paper, make two columns. Title the first "Control Group." Title the second "Experimental Group." Under the group names, draw a line for each participant you will need. Thus, if you were planning to use eight participants (four in each group), you would draw four lines under each group name.

CONTROL GROUP	EXPERIMENTAL GROUP
_____	_____
_____	_____
_____	_____
_____	_____

Step 2: Turn to a random numbers table, like the one at the end of this box (or the one in Appendix F). Roll a die to determine which column in the table you will use. Make a note in that column so that others could check your methods (Wilkinson & the Task Force on Statistical Inference, 1999).

Step 3: Assign the first number in the column to the first space under Control Group, the second number to the second space, and so on. When you have filled all the spaces for the control group, place the next number under the first space under Experimental Group and continue until you have filled all the spaces. Thus, if you used the random numbers table at the end of this box and you rolled a "5," you would start at the top of the fifth column of that table (the column starting with the number 81647), and your sheet of paper would look like this:

CONTROL GROUP	EXPERIMENTAL GROUP
81647	06121
30995	27756
76393	98872
07856	18876

Step 4: At the end of each control group score, write down a "C." At the end of each experimental group score, write down an "E." In this example, our sheet would now look like this:

CONTROL GROUP	EXPERIMENTAL GROUP
81647C	06121E
30995C	27756E
76393C	98872E
07856C	18876E

Step 5: Rank these numbers from lowest to highest. Then, on a second piece of paper, put the lowest number on the top line, the second lowest number on the next line, and so on. In this example, your page would look like this:

06121E	30995C
07856C	76393C
18876E	81647C
27756E	98872E

Step 6: Label the top line "Participant 1," the second line "Participant 2," and so forth. The first participant who shows up will be in the condition specified on the top line, the second participant who shows up will be in the condition specified by the second line, and so forth. In this example, the first participant will be in the experimental group, the second in the control group, the third and fourth in the experimental group, the fifth, sixth, and seventh in the control group, and the eighth in the experimental group. Thus, our sheet of paper would look like this:

Participant Number 1 = 06121E
Participant Number 2 = 07856C
Participant Number 3 = 18876E
Participant Number 4 = 27756E
Participant Number 5 = 30995C
Participant Number 6 = 76393C
Participant Number 7 = 81647C
Participant Number 8 = 98872E

(Continued)

BOX 10.1 Continued

Step 7: To avoid confusion, recopy your list, but make two changes. First, delete the random numbers. Second, write out "Experimental" and "Control." In this example, your recopied list would look like the following:

Participant Number 1 = Experimental
Participant Number 2 = Control
Participant Number 3 = Experimental
Participant Number 4 = Experimental
Participant Number 5 = Control
Participant Number 6 = Control
Participant Number 7 = Control
Participant Number 8 = Experimental

RANDOM NUMBERS TABLE

	COLUMN					
Row	**1**	**2**	**3**	**4**	**5**	**6**
1	10480	15011	01536	02011	81647	69179
2	22368	46573	25595	85393	30995	89198
3	24130	48360	22527	97265	76393	64809
4	42167	93093	06243	61680	07856	16376
5	37570	39975	81837	76656	06121	91782
6	77921	06907	11008	42751	27756	53498
7	99562	72905	56420	69994	98872	31016
8	96301	91977	05463	07972	18876	20922

will cause people to be happier, the null hypothesis would be getting 3 hours of full-spectrum lighting will have *no* demonstrated effect on happiness.

If your results show that the difference between groups is probably not due to chance, you can reject the null hypothesis. By rejecting the null hypothesis, you tentatively accept the experimental hypothesis: You conclude that the treatment has an effect.

But what happens if you fail to demonstrate conclusively that the treatment has an effect? Can you say that there is no effect for full-spectrum lighting? No, you can only say that you failed to prove beyond a reasonable doubt that full-spectrum lighting causes a change in happiness. In other words, you're back to where you were before you began the study: You do not know whether full-spectrum lighting causes a change in happiness.[2]

To reiterate a key point, *the failure to find a treatment effect doesn't mean that the treatment has no effect*. If you had looked more carefully, you might have found the effect.

To help yourself remember that you can't prove the null hypothesis, think of the null hypothesis as saying, "The difference between conditions *may* be due to chance." Even if you could prove that "The difference may be due to

[2]Those of you who are intimately familiar with confidence intervals may realize that null results do not necessarily send the researcher back to square one. Admittedly, we do not know whether the effect is greater than zero, but we could use confidence intervals to estimate a range in which the effect size probably lies. That is, before the study, we may have no idea of the potential size of the effect. We might think the effect would be anywhere between –100 units and +100 units. However, based on the data collected in the study, we could estimate, with 95% confidence, that the effect is between a certain range. For example, we might find, at the 95% level of confidence, that the effect is somewhere in the range between –1 units and +3 units.

chance," what would you have you proved? Certainly, you would not have proved that the difference *is* due to chance.

Conclusions About Experimental and Null Hypotheses

In summary, you have learned four important points about experimental and null hypotheses:

1. The experimental hypothesis is that the treatment has an effect.
2. The null hypothesis is that the treatment has no effect.
3. If you reject the null hypothesis, you can tentatively accept the hypothesis that the treatment has an effect.
4. If you fail to reject the null hypothesis, you can't draw any conclusions.

To remember these four key points, think about these hypotheses in the context of a criminal trial. In a trial, the *experimental* hypothesis is that the defendant *did* cause the crime; the *null* hypothesis is that the defendant did *not* commit the crime. The prosecutor tries to disprove the null hypothesis so that the jury will accept the experimental hypothesis. In other words, the prosecutor tries to disprove, beyond a reasonable doubt, the hypothesis that the defendant is "not guilty." If the jury decides that the null hypothesis is highly unlikely, they reject it and find the defendant guilty. If, on the other hand, they still have reasonable doubt, they fail to reject the null hypothesis and vote "not guilty." Note that their "not guilty" verdict is not an "innocent" verdict. Instead, it is a verdict reflecting that they are not sure, beyond a reasonable doubt, that the null hypothesis is false.

Manipulating the Independent Variable

Once you have your hypotheses, your next step is to manipulate the treatment. In any experiment, "participants are presented with the same general scenario (e.g., rating photographs of potential dating partners), but at least one aspect of this general scenario is manipulated" (Ickes, 2003, p. 22). In the simplest case of manipulating the treatment, you administer (assign) the treatment to some participants and withhold it from others. To isolate the treatment's effect, the conditions must be the same except for the treatment manipulation, as in the following classic experiments:

- In the first study showing that leading questions could bias eyewitness testimony, Loftus (1975) had students watch a film of a car accident and then gave students a questionnaire. The manipulation was whether the first question on the questionnaire was "How fast was Car A going when it ran the stop sign?"—a misleading question because Car A did *not* run the stop sign—*or* "How fast was Car A going when it turned right?"—a question that was not misleading.
- In the first study showing that people's entire impressions of another person could be greatly influenced by a single trait, Asch (1946) had participants think about a person who was described as either (a) "intelligent, skillful, industrious, *warm*, determined, practical, cautious" *or* (b) "intelligent, skillful, industrious, *cold*, determined, practical, cautious."
- In the first study showing that sex role stereotypes affect how people perceive infants, Condry and Condry (1976) had all participants use a form to rate the same baby. The only difference between how participants were

treated was whether the infant rating form listed the infant's name (a) as "David" and sex as "male" *or* (b) as "Dana" and sex as "female."

- In the first study showing that the pronouns people use when they describe their closest relationships affect how people see those relationships, Fitzsimons and Kay (2004) had all participants rate their relationship with their closest same-sex friend after writing five sentences about that friend. The only difference between groups was that one group was told to begin each sentence with "We," and was given the example, "We have known each other for 2 years," whereas the other group was told to begin each sentence with "(Insert friend's name) and I," and given the example, "John and I have known each other for 2 years."

To understand how you would manipulate a treatment, let's go back to trying to test the hypothesis about the effect of full-spectrum lighting on mood. To do this, you must vary the amount of light people get—and the amount should be independent of (should not depend on or be affected by) the individual's personal characteristics. To be specific, the amount of full-spectrum light participants receive should be determined by independent random assignment. Because the amount *varies* between the treatment group and the no-treatment group, because it varies *independently* of each participant's characteristics, and because it is determined by *independent* random assignment, full-spectrum lighting (the experimental intervention) is the **independent variable**.

In simple experiments, there are two values, or **levels of an independent variable**. The two levels can be types of treatment (e.g., lighting versus psychotherapy) or amounts (e.g., 1 hour of lighting versus 2 hours of lighting). In our lighting experiment, participants are randomly assigned to one of the following two levels of the independent variable: (1) 3 hours of full-spectrum lighting and (2) no full-spectrum lighting.

Experimental and Control Groups: Similar, but Treated Differently

The participants who are randomly assigned to get the higher level of the treatment (3 hours of full-spectrum light) are usually called the **experimental group**. The participants who are randomly assigned to get a lower level of the treatment (in this case, no treatment) are usually called the **control group**. Thus, in our example, the experimental group is the treatment group and the control group is the no-treatment group.

The control group is a comparison group. We compare the experimental (treatment) group to the control (no-treatment) group to see whether the treatment had an effect. If the treatment group scores the same as the comparison group, we would suspect that the treatment group would have scored that way even without the treatment. If, on the other hand, the treatment group scores differently than the control group, we would suspect that the treatment had an effect. For example, Ariely (2007) gave experimental group participants a chance to cheat. After taking a 50-item test, all participants transferred their answers from their tests to an answer sheet. For participants in the experimental group, the answer sheets already had the correct answers marked. Experimental group participants then shredded their tests and handed in their answer sheets. In this condition, students averaged about

36 questions correct. Did they cheat—and, if they did, how could Ariely possibly know? The only way to find out whether the experimental group cheated was to compare their scores to control group participants who were not allowed to cheat. Those control participants answered only about 33 questions correctly. By comparing the experimental group to the control group, Ariely found out that the experimental group cheated. Note that his conclusion—like that of any experimenter who uses a control group—only makes sense if the groups were equivalent at the start of the experiment. Thus, experimenters need to make sure that there are no systematic differences between the groups before the experimenter gives the groups different levels of the independent variable.

As the terms *experimental group* and *control group* imply, you should have several participants (preferably more than 30) in each of your conditions. The more participants you have, the more likely it is that your two groups will be similar at the start of the experiment. Conversely, the fewer participants you have, the less likely it is that your groups will be similar before you administer the treatment. For example, if you are doing an experiment to evaluate the effect of a strength pill and have only two participants (a 6 ft 4 in., 280-lb [1.9 m, 127 kg] offensive tackle and a 5 ft 1 in., 88-lb [1.5 m, 40 kg] person recovering from a long illness), random assignment will not have the opportunity to make your "groups" equivalent. Consequently, your control group would not be a fair comparison group.

The Value of Independence: Why Control and Experimental Groups Shouldn't Be Called "Groups"

Although we have noted that the experimental and control groups are groups in the sense that there should be several participants in each "group," that is the only sense in which these "groups" are groups. To conduct an experiment, you do *not* find two groups of participants and then randomly assign one group to be the experimental group and the other to be the control group.

Why You Should Not Choose Two Preexisting Groups

To see why not, suppose you were doing a study involving 10,000 janitors at a Los Angeles company and 10,000 managers at a New York company. You have 20,000 people in your experiment: one of the largest experiments in history. Then, you flip a coin and—on the basis of that single coin flip—assign the LA janitors to no treatment and the New York managers to treatment. Even though you have 10,000 participants in each group, your treatment and no-treatment groups differ in at least two systematic ways (where they live and what they do) before the study begins. Your random assignment is no more successful in making your groups similar than it was when you had only two participants. Consequently, to get random assignment to equalize your groups, you need to assign each participant ***independently***: individually, without regard to how previous participants were assigned.

Why You Should Not Let Your Groups Become "Groups"

Your concern with independence does not stop at assignment. After you have independently assigned participants to condition, you want each of your participants to remain independent. To maintain independence, do not test the

control participants in one group session and the experimental participants in a separate group session. Having one testing session for the control group and a second session for the experimental group hurts independence in two ways.

First, when participants are tested in groups, they may become group members who influence each other's responses rather than independent individuals. For example, instead of giving their own individual, independent responses, participants might respond as a conforming mob.

As a concrete example of the perils of letting participants interact, imagine that you are doing an ESP experiment. In the control group, only 30 of the 60 participants correctly guessed that the coin would turn up heads. In the experimental group, on the other hand, all 60 participants correctly guessed that the coin would turn up heads. Had each experimental group participant made his or her decision independently, such results would rarely[3] happen by chance. Thus, we would conclude that the treatment had an effect. However, if all the experimental group members talked to one another and made a group decision, they were not acting as 60 individual participants but as one group. In that case, the results would not be so impressive: Because all 60 experimental participants acted as one, the chances of all of them correctly guessing the coin flip were the same as the chances of one person correctly guessing a coin flip: 1 in 2 (50%).

Although this example shows what can happen when participants are tested in groups and allowed to interact freely, interaction can disturb independence even when group discussion is prohibited. Participants may influence one another through inadvertent outcries (laughs, exclamations like, "Oh no!") or through subtle nonverbal cues. In our lighting–happiness experiment, if we tested all the participants in a single group session, one participant who is crying uncontrollably might cause the entire experimental group to be unhappy, thereby leading us to falsely conclude that the lighting caused unhappiness. If, on the other hand, we tested each participant individually, the unhappy participant's behavior would not affect anyone else's responses.

The second reason for not testing all the experimental participants in one session and all the control participants in another is that such group testing turns the inevitable, random differences between testing sessions into systematic effects. For instance, suppose that when the experimental group was tested, there was a distraction in the hall, but there was no such distraction while the control group was tested. Like the treatment, this distraction was presented to all the experimental group participants, but to none of the control group participants. Thus, if the distraction did have an effect, its effect might be mistaken for a treatment effect. If, on the other hand, participants were tested individually, it is unlikely that only the experimental participants would be exposed to distractions. Instead, distractions would have a chance to even out so that participants in both groups would be almost equally affected by distractions.

But what if you are sure you won't have distractions? Even then, the sessions will differ in ways unrelated to the treatment. If you manage to test the participants at the same time, you'll have to use different experimenters and

---

[3] To be more precise, it should happen with a probability of $(1/2)^{60}$, which is less than .000000000000000009% of the time.

different testing rooms. If you manage to use the same experimenter and testing room, you'll have to test the groups at different times. Consequently, if you find a significant difference between your groups, you will have trouble interpreting those results. Specifically, you have to ask, "Is the significant difference due to the groups getting different levels of the treatment or to the groups being tested under different conditions (e.g., having different experimenters or being tested at different times of day)?"

To avoid these problems in interpreting your results, make sure that the treatment is the only factor that systematically varies. In other words, use independent random assignment and then test your participants individually (or in small groups) so that random differences between testing sessions have a chance to even out. If you must run participants in large groups, do not run groups made up exclusively of *either* experimental or control participants. Instead, run groups made up of *both* control and experimental participants.

The Value of Assignment (Manipulating the Treatment)

We have focused on the importance of independence to independent random assignment. Independence helps us start the experiment with two "groups" of participants that do not differ in any systematic way. But assignment is also a very important aspect of independent random assignment.

Random Assignment Makes the Treatment the Only Systematic Difference Between Groups

Random assignment to treatment group helps ensure that the only systematic difference between the groups is the treatment. With random assignment, our groups will be equivalent on the nontreatment variables we know about as well as on the (many) nontreatment variables we don't know about.

In our experiment, random assignment makes it so that one random sample of participants (the experimental group) is assigned to receive a high level of the independent variable whereas the other random sample of participants (the control group) is assigned to receive a low level of the independent variable. If, at the end of the study, the groups differed by more than would be expected by chance, we could say that the difference was due to the only nonchance difference between them: the treatment.

Without Random Assignment You Do Not Have a Simple Experiment

If you cannot randomly assign participants to your different groups, you cannot do a simple experiment. Because you cannot randomly assign participants to have certain personal characteristics, simple experiments cannot be used to study the effects of participant characteristics such as gender, race, personality, and intelligence.[4] For example, it makes no sense to assign a man to be a woman, a 7 ft 2 in. (218 cm) person to be short, or a shy person to be outgoing.

---

[4] You can, however, use experiments to investigate how participants react to people who vary in terms of these characteristics. For example, you can have an experiment in which participants read the same story except that one group is told that the story was written by a man, whereas the other group is told that the story was written by a woman. Similarly, you can randomly determine, for each participant, whether the participant interacts with a male or female experimenter.

To see why we need to be able to assign participants, let's imagine that you try to look at the effects of lighting on mood without using random assignment. Suppose you get a group of people who use light therapy and compare them to a group of people who do not use light therapy. What would be wrong with that?

The problem is that you are selecting two groups of people who you know are different in at least one way, and then you are assuming that they don't differ in any other respect. The assumption that the groups are identical in every other respect is probably wrong. The light therapy group probably feels more depressed, lives in colder climates, is more receptive to new ideas, and is richer than the other group.

Because the groups differ in many ways other than in terms of the "treatment," it would be foolish to say that the treatment—rather than one of these many other differences between the groups—is what caused the groups to score differently on the happiness measure. For example, if the group of light users is more depressed than our sample of nonusers, we could not conclude that the lighting caused their depression. After all, the lighting might be a partial cure for—rather than a cause of—their depression.

But what if the group of lighting users is less depressed? Even then, we could not conclude that the lighting is causing an effect. Lighting users may be less depressed because they are richer, have more spare time, or differ in some other way from those who don't use lights. In short, if you do not randomly *assign* participants to groups, you cannot conclude anything about the effects of a treatment.

If, on the other hand, you start with one group of participants and then randomly assign half to full-spectrum lighting and half to normal lighting, interpreting differences between the groups would be much simpler. Because the groups probably were similar before the treatment was introduced, large group differences in happiness are probably due to the only systematic difference between them—the lighting.

Collecting the Dependent Variable

Before you can determine whether the lighting caused the experimental group to be happier than the control group, you must measure each participant's happiness. You know that each person's happiness will be somewhat *dependent* on the individual's personality and you predict that his or her score on the happiness *variable* will also be *dependent* on the lighting. Therefore, scores on the happiness measure are your **dependent variable**. Because the dependent variable is what the participant does that you *measure*, the dependent variable is also called the **dependent measure**.

The Statistical Significance Decision: Deciding Whether to Declare That a Difference Is Not a Coincidence

After measuring the dependent variable, you will want to compare the experimental group's happiness scores to the control group's. One way to make this comparison is to subtract the average of the happiness scores for the control (comparison) group from the average of the experimental group's happiness scores.

Unfortunately, knowing how much the groups differ doesn't tell you how much of an effect the treatment had. After all, even if the treatment had no

effect, nontreatment factors would probably still make the groups differ. In other words, even if the treatment had no effect, the groups may differ due to random error.

How can you determine that the difference between groups is due to something more than random error? To determine the probability that the difference is not exclusively due to chance, you need to use **inferential statistics**: the science of chance.

Statistically Significant Results: Declaring That the Treatment Has an Effect

If, after using statistics, you find that the difference between your groups is greater than could be expected if only chance were at work, your results are statistically significant. The term **statistical significance** means that you are sure, beyond a reasonable doubt, that the difference you observed is not a fluke.

What is a reasonable doubt? Usually, before researchers commit themselves to saying that the treatment has an effect, they want a 5% probability ($p = .05$) or less ($p < .05$) that they would get such a pattern of results when there really was no effect. Consequently, you may hear researchers say that their results were "significant at the point-oh-five level" and, in journal articles, you will often see statements like, "the results were statistically significant ($p < .05$)."

To review, if you do a simple experiment, you will probably find that the treatment group mean is different from the control group mean. Such a difference is not, by itself, evidence of the treatment's effect. Indeed, because random assignment does not create identical groups, you would expect the two group means to differ to some extent. Therefore, the question is not "Is there a difference between the group means?" but rather "Is the difference between the group means a reliable one—one bigger than would be expected if only random factors were at work?" To answer that question, you need to use statistics.

By using statistics, you might find that if only chance factors were at work (i.e., if the independent variable had no effect), you would get a difference as large as that less than 5% of the time. If differences as big or bigger than what you found occur less than 5% of the time by chance alone ($p < .05$) when the null hypothesis is true, you would probably conclude that the null hypothesis is not true. To state your conclusion more formally, you might say that "the results are statistically significant at the .05 level." By "statistically significant," you mean that because it's unlikely that the difference between your groups is due to chance alone, you conclude that some of the difference was due to the treatment. With statistically significant results, you would be relatively confident that if you repeated the study, you would get the same pattern of results—the independent variable would again cause a similar type of change in the scores on the dependent variable. In short, statistical significance suggests that the results are reliable and replicable.

Statistically Significant Effects May Be Small

Statistical significance, however, does not mean that the results are significant in the sense of being large. Just because a difference is statistically significant—reliably different from zero—doesn't mean the difference is large. Even a tiny difference can be statistically reliable. For example, if you flipped a coin 5,000

times and it came up heads 51% of the time, this 1% difference from what would be expected by chance (50% heads) would be statistically significant.

Statistically Significant Results May Be Insignificant (Trivial)

Nor does statistical significance mean that the results are significant in the sense of being important. If you have a meaningless hypothesis, you may have results that are statistically significant but scientifically meaningless.

Statistically Significant Results May Refute Your Experimental Hypothesis

Finally, statistically significant results do not necessarily support your hypothesis. For example, suppose your hypothesis is that the treatment improves behavior. A statistically significant effect for the treatment would mean that the treatment had an effect. But did the treatment improve behavior or make it worse? To find out, you have to look at the means to see whether the treatment group or no-treatment group is behaving better.

Summary of the Limitations of Statistically Significant Results

In short, statistically significant results tell you nothing about the direction, size, or importance of the treatment effect (see Table 10.1). Because of the limitations of statistical significance, the American Psychological Association appointed a task force to determine whether significance testing should be eliminated. The task force did "not support any action that could be interpreted as banning the use of null significance testing or *p* values in psychological research and publication" (American Psychological Association, 1996b). However, the task force did recommend that, in addition to reporting whether the results were statistically significant, authors should provide information about the direction and size of effects.

Null Results: Why We Can't Draw Conclusions From Nonsignificant Results

You now know how to interpret statistically significant results. But what if your results are *not* statistically significant? That is, what if you can't reject the null hypothesis that the difference between your groups could be due to chance? Then, you have *failed* to reject the null hypothesis; therefore, your results would be described as "not significant."

As the phrase "not significant" suggests, you can't draw any conclusions from such findings. With **nonsignificant results** (also called **null results**), you

TABLE **10.1**
Limits of Statistical Significance

Statistically significant differences are

1. probably not due to chance alone
2. not necessarily large
3. not necessarily in the direction you predicted
4. not necessarily important

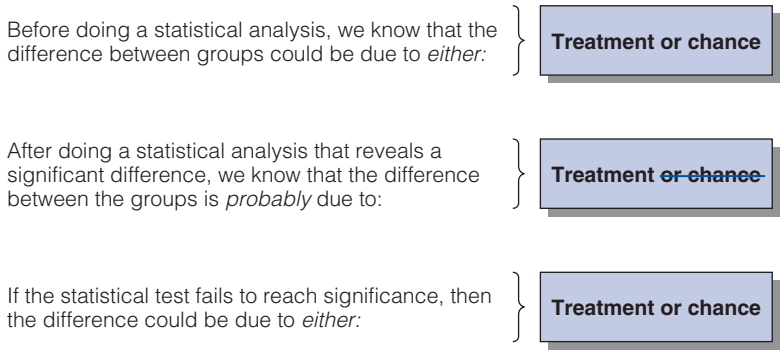

FIGURE 10.1 The Meaning of Statistical Significance

If the results are statistically significant, we can conclude that the difference between the groups is not due entirely to chance and therefore some of the difference must be due to the treatment. However, if the results are not statistically significant, the results could be due to chance or treatment. Put another way, we don't know any more than we did before we subjected the results to statistical analysis.

TABLE 10.2
Common Errors in Discussing Null Results

STATEMENT	FLAW
"The results were not significant. Therefore, the independent variable had no effect."	"Not that I know of" is not the same as proving "there isn't any."
"The treatment had an effect, even though the results are not significant."	"Not significant" means that you failed to find an effect. Therefore, the statement could be translated as, "I didn't find an effect for the treatment, but I really did."

do not know whether the treatment has an effect that you failed to find or whether the treatment really has no effect (see Figure 10.1).

Nonsignificant results are analogous to a "not guilty" verdict: Is the defendant innocent, or did the prosecutor present a poor case? Often, defendants get off, not because of overwhelming proof of their innocence, but because of lack of conclusive proof of their guilt.

You have seen that nonsignificant results neither confirm nor deny that the treatment had an effect. Unfortunately, you will find some incompetents treating null results as proof that the treatment has an effect—whereas other bad researchers will treat null results as proof that the treatment has no effect (see Table 10.2).

Nonsignificant Results Are Not Significant

All too often, people act like nonsignificant results are really significant. They may say, "The difference between my groups shows that the treatment had an

effect, even though the difference is not significant." Reread the previous quote because you're sure to see it again: It's one of the most common contradictory statements that researchers make. People making this statement are really saying, "The difference is due to the treatment, even though I've found no evidence that the difference isn't simply due to chance."

Null Results Do Not Prove the Null Hypothesis: "I Didn't Find It" Doesn't Mean It Doesn't Exist

As we have just discussed, some people act like null results secretly prove the experimental hypothesis. On the other hand, some people make the opposite mistake: They incorrectly assume that null results prove the null hypothesis. That is, they falsely conclude that null results prove that the treatment had no effect. Some individuals make this mistake because they think the term "null results" implies that the results prove the null hypothesis. Those people would be better off thinking of null results as "no results" than to think that null results support the null hypothesis.

Thinking that nonsignificant results support the null hypothesis is a mistake because it overlooks the difficulty of conclusively proving that a treatment has an effect. People should realize that not finding something is not the same as proving that the thing does not exist. After all, people often fail to find things that clearly exist, such as books that are in the library, items that are in the grocery store, and keys that are on the table in front of them.

Even in highly systematic investigations, failing to find something doesn't mean the thing does not exist. For example, in 70% of all murder investigations, investigators do not find a single identifiable print at the murder scene—not even the victim's. Thus, the failure to find the suspect's fingerprints at the scene is hardly proof that the suspect is innocent. For essentially the same reasons, the failure to find an effect is not proof that there is no effect.

Summary of the "Ideal" Simple Experiment

Thus far, we have said that the simple experiment gives you an easy way to determine whether a factor causes an effect. If you can randomly assign participants to either a treatment or no-treatment group, all you have to do is find out whether your results are statistically significant. If your results are statistically significant, your treatment probably had an effect. No method allows you to account for the effects of nontreatment variables with as little effort as random assignment.

ERRORS IN DETERMINING WHETHER RESULTS ARE STATISTICALLY SIGNIFICANT

There is one drawback to random assignment: Differences between groups may be due to chance rather than to the treatment. Admittedly, statistical tests—by allowing you to predict the extent to which chance may cause the groups to differ—minimize this drawback. Statistical tests, however, do not allow you to perfectly predict chance all of the time. Therefore, you may err by either underestimating or overestimating the extent to which chance is causing your groups to differ (see Table 10.3).

TABLE **10.3**
Possible Outcomes of Statistical Significance Decision

	REAL STATE OF AFFAIRS	
STATISTICAL SIGNIFICANCE DECISION	**Treatment has an effect**	**Treatment does not have an effect**
Significant: Reject the null hypothesis	Correct decision	Type 1 error
Not significant: Do not reject the null hypothesis	Type 2 error	Correct decision

Type 1 Errors: "Crying Wolf"

If you underestimate the role of chance, you may make a **Type 1 error**: mistaking a chance difference for a real difference. In the simple experiment, you would make a Type 1 error if you mistook a chance difference between your experimental and control groups for a treatment effect. More specifically, you would make a Type 1 error if you declared that a difference between your groups was statistically significant, when the treatment really didn't have an effect. In nonresearch settings, examples of Type 1 errors include:

- a jury convicting an innocent person because they mistake a series of coincidences as evidence of guilt
- a person responding to a false alarm, such as thinking that the phone is ringing when it's not or thinking that an alarm is going off when it's not
- a physician making a "false positive" medical diagnosis, such as telling a woman she is pregnant when she isn't

Reducing the Risk of a Type 1 Error

What can you do about Type 1 errors? *There is only one thing you can do: You can decide what risk of a Type 1 error you are willing to take.* Usually, experimenters decide that they are going to take less than a 5% risk of making a Type 1 error. In other words, they say their results must be significant at the .05 level ($p < .05$) before they declare that their results are significant. They are comfortable with the odds of their making a Type 1 error being less than 5 in 100. But why take even that risk? Why not take less than a 1% risk?

Accepting the Risk of a Type 1 Error

To understand why not, imagine you are betting with someone who is flipping a coin. For all 10 flips, she calls "heads." She wins most of the 10 flips.

Let's suppose that you will refuse to pay up if you have statistical proof that she is cheating. However, you do not want to make the Type 1 error of attributing her results to cheating (using a biased coin) when the results are due only to luck. How many of the 10 flips does she have to win before you "prove" that she is cheating?

To help you answer this question, we looked up the odds of getting 8, 9, or 10 heads in 10 flips of a fair coin.[5] Those odds are as follows:

EVENT	PROBABILITY EXPRESSED IN PERCENTAGES	PROBABILITY EXPRESSED IN DECIMAL FORM
Chances of 8 or more heads	5.47%	.0547
Chances of 9 or more heads	1.08%	.0108
Chances of 10 heads	0.1%	.001

From these odds, you can see that you can't have complete, absolute proof that she is cheating. Thus, if you insist on taking 0% risk of falsely accusing her (you want to be absolutely 100% sure), you would not call her a cheat—even if she got 10 heads in a row. As you can see from the odds we listed, it is very unlikely (.1% chance), but still possible, that she could get 10 heads in a row, purely by chance alone. Consequently, if you are going to accuse her of cheating, you are going to have to take some risk of making a false accusation.

If you were willing to take more than a 0% risk but were unwilling to take even a 1% risk of falsely accusing her (you wanted to be more than 99% sure), you would call her a cheat if all 10 flips turned up heads—but not if 9 of the flips were heads. If you were willing to take a 2% risk of falsely accusing her (you wanted to be 98% sure), you would call her a cheat if either 9 or 10 of the flips turned up heads. Finally, if you were willing to take a 6% risk of falsely accusing her (you would settle for being 94% sure), you could refuse to pay up if she got 8 or more heads.

This betting example gives you a clue about what happens when you set your risk of making a Type 1 error. When you determine your risk of making a Type 1 error, you are indirectly determining how much the groups must differ before you will declare that difference statistically significant. If you are willing to take a relatively large risk of mistaking a difference that is due only to chance for a treatment effect, you may declare a relatively small difference statistically significant. If, on the other hand, you are willing to take only a tiny risk of mistakenly declaring a chance difference statistically significant, you must require that the difference between groups be relatively large before you are willing to call it statistically significant. In other words, all other things being equal, the larger the difference must be before you declare it significant, the less likely it is that you will make a Type 1 error. To take an extreme example of this principle, if you would not declare even the biggest possible difference between your groups statistically significant, you would never make a Type 1 error.

[5] You do not need to know how to calculate these percentages.

Type 2 Errors: "Failing to Announce the Wolf"

The problem with not taking any risk of making a Type 1 error is that, if the treatment did have an effect, you would be unable to detect it. In trying to be very sure that a difference is due to treatment and not to chance, you may make a **Type 2 error**: overlooking a genuine treatment effect because you think the differences between conditions might be due to chance. Examples of Type 2 errors in nonresearch situations include:

- a jury letting a criminal go free because they wanted to be sure beyond any doubt and they realized that it was possible that the evidence against the defendant was due to numerous, unlikely coincidences
- a person failing to hear the phone ring
- a radar detector failing to detect a speed trap
- a physician making a "false negative" medical diagnosis, such as failing to detect that a woman was pregnant

In short, whereas Type 1 errors are errors of commission (yelling "fire" when there is no fire), Type 2 errors are errors of omission (failing to yell "fire" when there is a fire). In trying to avoid Type 1 errors, you may increase your risk of making Type 2 errors. In the extreme case, if you were never willing to risk making a Type 1 error, you would never detect real treatment effects. But because you want to detect real treatment effects, you will take a risk of making a Type 1 error—and you will take steps to improve your study's **power**: the ability to find real differences and declare those differences statistically significant; or, put another way, the ability to avoid making Type 2 errors.[6]

The Need to Prevent Type 2 Errors: Why You Want the Power to Find Significant Differences

You can have power without increasing your risk of making a Type 1 error. Unfortunately, many people don't do what it takes to have power.

If you don't do what it takes to have power, your study may be doomed: Even if your treatment has an effect, you will fail to find that effect statistically significant. In a way, looking for a significant difference between your groups with an underpowered experiment is like looking for differences between cells with an underpowered microscope.

As you might imagine, conducting a low-powered experiment often leads to frustration over not finding anything. Beginning researchers frequently frustrate themselves by conducting such low-powered experiments. (We

---

[6] In a sense, power (defined as 1.00 – the probability of making a Type 2 error) and Type 2 errors are opposites. *Power* refers to the chances (given that the treatment really does have a certain effect) of *finding* a significant treatment effect, whereas the probability of a *Type 2 error* refers to the chances (given that the treatment really does have a certain effect) of *failing to find* a significant treatment effect. If you plug numbers into the formula "1.00 – power = chances of making a Type 2 error," you can see that power and Type 2 errors are inversely related. For example, if power is 1, you have a *0%* chance of making a Type 2 error (because 1.00 – *1.00* = 0%). Conversely, if the treatment has an effect and power is 0, you have a *100%* chance of making a Type 2 error (because 1.00 – *0* = 100%). Often, power is around .40, meaning that, if the treatment has an effect, the researcher has a 40% (.40) chance of finding that effect and a *60%* chance of not finding that effect (because 1.00 – *.40* = 60%).

know we did.) Why do beginning researchers often fail to design sufficiently powerful experiments?

STATISTICS AND THE DESIGN OF THE SIMPLE EXPERIMENT

One reason inexperienced researchers fail to design powerful experiments is they simply do not think about power—a "sin" that many professional researchers also commit (Cohen, 1990). But even when novice researchers do think about power, they often think that it is a statistical concept and therefore has nothing to do with design of experiments. Admittedly, power is a statistical concept. However, *statistical concepts should influence the design of research*. Just as a bridge builder should consider engineering principles when designing a bridge, a researcher should consider statistical principles when designing a study. If you consider statistical power when designing your study, your study should have enough power to find the differences that you are looking for—if those differences really exist.

Power and the Design of the Simple Experiment

To have enough power, you must reduce the risk of chance differences hiding the treatment effect. As you can see from Figure 10.2, two ways to stop random error from overwhelming your treatment effect are (1) reduce the effects of random error and (2) increase the size of the treatment effect.

Reduce the Effect of Random Error

One of the most obvious ways to reduce the effects of random error is to reduce the potential sources of random error. The major sources of random error are random differences between testing situations, random measurement error, random differences between participants, and sloppy coding of data.

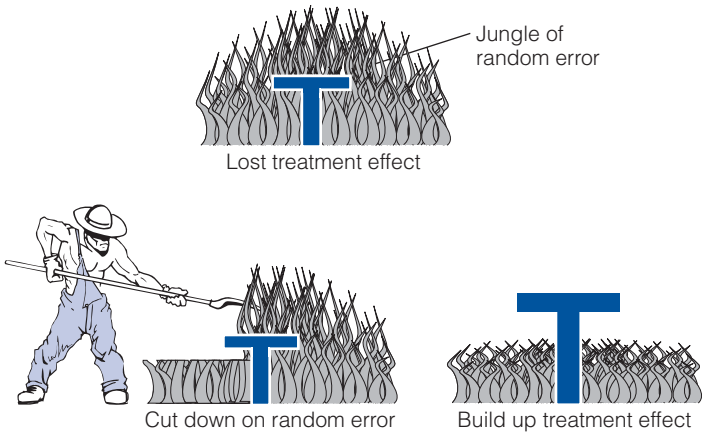

FIGURE **10.2** Cutting Down on Random Error and Building Up the Treatment Effect: Two Ways to Avoid Losing Your Treatment Effect in a "Jungle" of Random Error

Standardize Procedures and Use Reliable Measures. Because a major source of random error is random variation in the testing situation, you can reduce random error by standardizing your experiment. Standardization consists of keeping the testing environment and the experimental procedures as constant as possible. Thus, to improve power, you might want the noise level, illumination level, temperature, and other conditions of testing to be the same for each participant. Furthermore, you would want to treat all your experimental group participants identically and treat all your control group participants identically. In addition to reducing random error by standardizing procedures, you should also reduce random error by using a reliable dependent measure (for more about how reliable measures boost power, see Chapter 6).

The desire for both reliable measures and strict standardization makes some psychologists love both instruments and the laboratory. Under the lab's carefully regulated conditions, experimenters can create powerful and sensitive experiments.

Other experimenters, however, reject the laboratory setting in favor of real-world settings. By using real-world settings, they can more easily make a case for their study's external validity. The price they pay for leaving the laboratory is that they are no longer able to keep many nontreatment variables (temperature, distractions, noise level, etc.) constant. These variables, free to vary wildly, create a jungle of random error that may hide the treatment's effect.

Because of the large variability in real-world settings and the difficulties of using sensitive measures in the field, even die-hard field experimenters may first look for a treatment's effect in the lab. Only after they have found that the treatment has an effect in the lab will they try to detect the treatment's effect in the field.

Use a Homogeneous Group of Participants. Like differences between testing sessions, differences between participants can hide treatment effects. Even if the treatment effect causes a large difference between your groups, you may overlook that effect, mistakenly believing that the difference between your groups is due to your participants being years apart in age and worlds apart in terms of their experiences.

To decrease the chances that between-subject differences will mask the treatment's effect, choose participants who are similar to one another. For instance, select participants who are the same gender, same age, and have the same IQ—or, study rats instead of humans. With rats, you can select participants that have grown up in the same environment, have similar genes, and even have the same birthday. By studying homogeneous participants under standardized situations, rat researchers can detect very subtle treatment effects.

Code Data Carefully. Obviously, sloppy coding of the data can sabotage the most sensitively designed study. So, why do we mention this obvious fact?

We mention it because careful coding is a cheap way to increase power. If you increase power by using nonhuman animals as participants, you may lose the ability to generalize to humans. If you increase power by using a lab experiment rather than a field experiment, you may lose some of your ability

to generalize to real-world settings. But careful coding costs you nothing—except for a little time spent rechecking the coding of your data.

Let Random Error Balance Out. Thus far, we have talked about reducing the effects of random error by reducing the amount of random error. But you can reduce the *effects* of random error on your data without reducing the *amount* of random error in your data.

The key is to give random error more chances to balance out. To remind yourself that chance does balance out in the long run, imagine flipping a fair coin. If you flipped it six times, you might get five tails and one head—five times as many tails as heads. However, if you flipped it 1,000 times, you would end up with almost as many heads as tails.

Similarly, if you use five participants in each group, your groups probably won't be equivalent before the experiment begins. Thus, even if you found large differences between the groups at the end of the study, you might have to say that the differences could be due to chance alone. However, if you use 60 participants in each group, your groups should be equivalent before the study begins. Consequently, a treatment effect that would be undetected if you used 5 participants per group might be statistically significant if you used 60 participants per group. In short, to take advantage of the fact that random error balances out, boost your study's power by studying more participants.

Create Larger Effects: Bigger Effects Are Easier to See

Until now, we have talked about increasing power by making our experiment more sensitive to small differences. Specifically, we have talked about two ways of preventing the "noise" caused by random error from making us unable to "hear" the treatment effect: (1) reducing the amount of random error and (2) giving random error a chance to balance out. However, we have left out one obvious way to increase our experiment's ability to detect the effect: making the effect louder (bigger) and thus easier to hear.

As you might imagine, bigger effects are easier to find. But how do we create bigger effects? Your best bet for increasing the size of the effect is to give the control group participants a very low level of the independent variable while giving the experimental group a very high level of the independent variable. Hence, to have adequate power in the lighting experiment, rather than giving the control group 1 hour of full-spectrum light and the experimental group 2 hours, you might give the control group no full-spectrum light and the experimental group 4 hours of full-spectrum light.

To see how researchers can maximize the chances of finding an effect by giving the experimental and control groups widely different levels of treatment, let's consider an experiment by T. D. Wilson and Schooler (1991). Wilson and Schooler wanted to determine whether thinking about the advantages and disadvantages of a choice could hurt one's ability to make the right choice. In one experiment, they had participants rate their preference for the taste of several fruit-flavored jams. Half the participants rated their preferences after completing a "filler" questionnaire asking them to list reasons why they chose their major. The other half rated their preferences after completing a questionnaire asking them to "analyze why you feel the way you do about each jam in order to prepare yourself for your evaluations." As Wilson and Schooler predicted, the participants who thought about why they liked

the jam made less accurate ratings (ratings that differed more from experts' ratings) than those who did not think about why they liked the jam.

Although the finding that one can think too much about a choice is intriguing, we want to emphasize another aspect of Wilson and Schooler's study: the difference between the amount of time experimental participants reflected on jams versus the amount of time that control participants reflected on jams. Note that the researchers did not ask the control group to do any reflection whatsoever about the jams. To reiterate, Wilson and Schooler did not have the control group do a moderate amount of reflection and the experimental group do slightly more reflection. If they had, Wilson and Schooler might have failed to find a statistically significant effect.

Conclusions About How Statistical Considerations Impact Design Decisions

By now, you can probably appreciate why R. A. Fisher said, "To consult a statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of." The reason you should think about statistics before you do an experiment is that statistical considerations influence virtually every aspect of the design process (see Table 10.4). For example, statistical considerations even dictate what kind of hypothesis you can test. Because you cannot accept the null hypothesis, the only hypotheses that you can hope to support are hypotheses that the groups will differ. Therefore, you cannot do a simple experiment to prove that two treatments have the same effect or that a certain treatment will be just as ineffective as no treatment.

Not only do statistical considerations dictate what types of hypotheses you can have, but they also mandate how you should assign your participants.

TABLE **10.4**
Implications of Statistics for the Simple Experiment

STATISTICAL CONCERN/REQUIREMENT	IMPLICATIONS FOR DESIGNING THE SIMPLE EXPERIMENT
Observations must be independent.	You must use independent random assignment and, ideally, you will test participants individually.
Groups must differ for only two reasons—random differences and the independent variable.	You must randomly assign participants to groups.
It is impossible to accept the null hypothesis.	You cannot use the experiment to prove that a treatment has no effect or to prove that two treatments have identical effects.
You need enough power to find a significant effect.	You should 1. Standardize procedures. 2. Use sensitive, reliable dependent variables. 3. Code data carefully. 4. Use homogeneous participants. 5. Use many participants. 6. Use extreme levels of the independent variable.

Specifically, if you do not assign your participants to groups using independent random assignment, you do not have a valid experiment.

Statistical considerations also dictate how you should treat your participants. You will not have a valid experiment if you let participants influence one another's responses or if you do anything else that would violate the statistical requirement that individual participants' responses must be independent.

Even when statistics are not dictating what you must do, they are suggesting what you should do. To avoid making Type 2 errors, you should do the following:

1. Standardize your procedures.
2. Use sensitive and reliable dependent measures.
3. Carefully code your data.
4. Use homogeneous participants.
5. Use many participants.
6. Use extreme levels of the independent variable.

NONSTATISTICAL CONSIDERATIONS AND THE DESIGN OF THE SIMPLE EXPERIMENT

Statistical issues are not the only issues that you should consider when designing a simple experiment. If you considered only statistical power, you could harm your participants, as well as your experiment's external and construct validity. Therefore, in addition to statistical issues such as power, you must also consider external validity, construct validity, and ethical issues.

External Validity Versus Power

Many of the things you can do to improve your study's power may hurt your study's external validity. For example, using a laboratory setting, homogeneous participants, and extreme levels of the independent variable all improve power, but all may reduce external validity.

By using a lab experiment to stop unwanted variables from varying, you may have more power to find an effect. However, by preventing unwanted variables from varying, you may hurt your ability to generalize your results to real life—where these unwanted variables *do* vary.

By using a homogeneous set of participants (18-year-old, White males with IQs between 120 and 125), you reduce between-subject differences, thereby enhancing your ability to find treatment effects. However, because you used such a restricted sample, you would not be as able to generalize your results to the average American as a researcher whose participants were a random sample of Americans.

Finally, by using extreme levels of the independent variable, you may be able to find a significant effect for your independent variable. If you use extreme levels, though, you may be like the person who used a sledgehammer to determine the effects of hammers—you don't know the effect of realistic, naturally occurring levels of the treatment variable.

Construct Validity Versus Power

Your efforts to improve power may hurt not only your experiment's external validity but also its construct validity. For example, suppose you had two choices for your measure. The first is a 100-point rating scale that is sensitive and reliable. However, the measure is vulnerable to subject bias: If participants guess your hypothesis, they can easily circle the rating they think you want them to. The second is a measure that is not very reliable or sensitive, but it is a measure that participants couldn't easily fake. If power was your only concern, you would pick the first measure despite its vulnerability to subject bias. With it, you are more likely to find a statistically significant effect. However, because construct validity should be an important concern, many researchers would suggest that you pick the second measure.

If you sought only statistical power, you might also compromise the construct validity of your independent variable manipulation. For instance, to maximize your chances of getting a significant effect for full-spectrum lighting, you would give the experimental group full-spectrum lighting and make the control group an **empty control group**: a group that gets no kind of treatment. Compared to the empty control group, the treatment group

1. receives a gift (the lights) from the experimenter
2. gets more interaction with, and attention from, the experimenter (as the experimenter checks participants to make sure they are using the lights)
3. adopts more of a routine than the controls (using the lights every morning from 6:00 a.m. to 8:00 a.m.)
4. has higher expectations of getting better (because they have more of a sense of being helped) than the controls

As a result of all these differences, you would have a good chance of finding a significant difference between the two groups. Unfortunately, if you find a significant effect, it's hard to say that the effect is due to the full-spectrum lighting and not due to any of these other side effects of your manipulation.[7]

To minimize these side effects of the treatment manipulation, you might give your control group a **placebo treatment**: a substance or treatment that has no effect. Thus, rather than using a no-light condition, you might expose the control group to light from an ordinary 75-watt incandescent light bulb. You would further reduce the chances of bias if you made both the experimenters and participants **blind (masked)**: unaware of which kind of treatment the participant was getting. If you make the researcher who interacts with the participants blind, that researcher will not bias the results in favor of the experimental hypothesis. Similarly, by making participants blind, you make it less likely that participants will bias the results in favor of the hypothesis.

In short, the use of placebos, the use of **single blinds** (in which either the participant or the experimenter is blind), and the use of **double blinds** (in which both the participant and the experimenter are blind) all may reduce the chances that you will obtain a significant effect. However, if you use

[7] The problem of using an empty control group is even more apparent in research on the effect of surgery. For example, if a researcher finds that rats receiving brain surgery run a maze slower than a group of rats not receiving an operation, the researcher should not conclude that the surgery's effect was due to removing a part of the brain that plays a role in maze-running.

these procedures and still find a significant effect, you can be relatively confident that the treatment itself—rather than some side effect of the treatment manipulation—is causing the effect.

You have seen that what is good for power may harm construct validity, and vice versa. But what trade-offs should you make? To make that decision, you might find it helpful to see what trade-offs professional experimenters make between power and construct validity. Do experienced experimenters use empty control groups to get significant effects? Or, do they avoid empty control groups to improve their construct validity? Do they avoid blind procedures to improve power? Or, do they use blind procedures to improve construct validity?

Often, experimenters decide to sacrifice power for construct validity. For example, in their jam experiment, Wilson and Schooler did not have an empty control group. In other words, their control group did not simply sit around doing nothing while the experimental group filled out the questionnaire analyzing reasons for liking a jam. Instead, the control group also completed a questionnaire. The questionnaire was a "filler questionnaire" about their reasons for choosing a major. If Wilson and Schooler had used an empty control group, critics could have argued that it was the act of filling out a questionnaire—not the act of reflection—that caused the treatment group to make less accurate ratings than the controls. For example, critics could have argued that the controls' memory for the jams was fresher because they were not distracted by the task of filling out a questionnaire.

To prevent critics from arguing that the experimenters influenced participants' ratings, Wilson and Schooler made the experimenters blind. To implement the blind technique, Wilson and Schooler employed two experimenters. The first experimenter had participants (a) taste the jams and (b) fill out either the control group (filler) questionnaire or the experimental group (reasons) questionnaire. After introducing the participants to Experimenter 2, Experimenter 1 left the room. Then, Experimenter 2—who was unaware of (blind to) whether the participants had filled out the reasons or the filler questionnaire—had participants rate the quality of the jams.

Ethics Versus Power

As you have seen, increasing a study's power may conflict with both external and construct validity. In addition, increasing power may conflict with ethical considerations. For example, suppose you want to use extreme levels of the independent variable (food deprivation) to ensure large differences in the motivation of your animals. In that case, you need to weigh the benefits of having a powerful manipulation against ethical concerns, such as the comfort and health of your subjects (for more about ethical concerns, see Chapter 2 and Appendix D).

Ethical concerns determine not only how you treat the experimental group but also how you treat the control group. Just as it might be unethical to administer a potentially harmful stimulus to your experimental participants, it also might be unethical to withhold a potentially helpful treatment from your control participants. For instance, it might be ethically questionable to withhold a possible cure for depression from your controls. Therefore, rather than maximizing power by completely depriving the control group of a treatment, ethical concerns may dictate that you give the control group a

TABLE **10.5**
Conflicts Between Power and Other Research Goals

ACTION TO HELP POWER	HOW ACTION MIGHT HARM OTHER GOALS
Use a homogeneous group of participants to reduce random error due to participants.	May hurt your ability to generalize to other groups.
Test participants under controlled laboratory conditions to reduce the effects of extraneous variables.	1. May hurt your ability to generalize to real-life situations where extraneous variables are present. 2. Artificiality *may* hurt construct validity. If the setting is so artificial that participants are constantly aware that what they are doing is not real and just an experiment, they may *act* to please the experimenter rather than expressing their true reactions to the treatment.
Use artificially high or low levels of the independent variables to get big differences between groups.	1. You may be unable to generalize to realistic levels of the independent variable. 2. May be unethical.
Use an empty control group to maximize the chance of getting a significant difference between the groups.	Construct validity is threatened because the significant difference may be due to the participants' expectations rather than to the independent variable.
Test many participants to balance out the effects of random error.	Expensive and time-consuming.

moderate dose of the treatment. (For a summary of the conflicts between power and other goals, see Table 10.5.)

ANALYZING DATA FROM THE SIMPLE EXPERIMENT: BASIC LOGIC

After carefully weighing both statistical and nonstatistical considerations, you should be able to design a simple experiment that would test your experimental hypothesis in an ethical and internally valid manner. If, after consulting with your professor, you conduct that experiment, you will have data to analyze.

To understand how you are going to analyze your data, remember why you did the simple experiment. You did it to find out whether the treatment would have an effect on a unique population—all the participants who took part in your experiment. More specifically, you wanted to know the answer to the hypothetical question: "If I had put all my participants in the experimental condition, would they have scored differently than if I had put all of them in the control condition?" To answer this question, you need to know the averages of two **populations**:

> Average of Population #1—what the average score on the dependent measure would have been if all your participants had been in the control group.

Average of Population #2—what the average score on the dependent measure would have been if all your participants had been in the experimental group.

Unfortunately, you cannot measure both of these populations. If you put all your participants in the control condition, you won't know how they would have scored in the experimental condition. If, on the other hand, you put all your participants in the experimental condition, you won't know how they would have scored in the control condition.

Estimating What You Want to Know: Your Means Are Sample Means

You can't directly get the population averages you want, so you do the next best thing—you estimate them. You can estimate them because, thanks to independent random assignment, you split all your participants (your population of participants) into two random samples. That is, you started the experiment with two random samples from your original population of participants. These two "chips off the same block" were the control group and the experimental group (see Figure 10.3).

The average score of the random sample of your participants who received the treatment (the experimental group) is an estimate of what the average score would have been if all your participants received the treatment. The average score of the random sample of participants who received no treatment (the control group) is an estimate of what the average score would have been if all of your participants had been in the control condition.

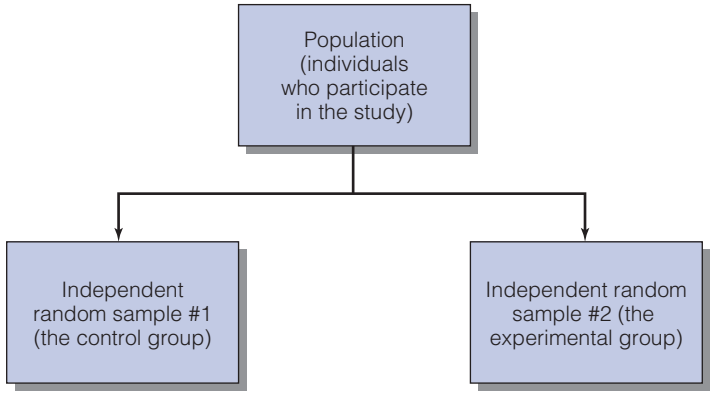

FIGURE **10.3** The Control Group and the Experimental Group Are Two Samples Drawn From the Same Population

Problem: If the average score for the experimental group is different from the average score for the control group, is this difference due to the two groups receiving different treatments? To random error related to sampling? (Two random samples from the same population may differ.)

Calculating Sample Means: Getting Your Estimates

Even though only half your participants were in the experimental group, you will assume that the experimental group is a fair sample of your entire population of participants. Thus, the experimental group's average score should be a good estimate of what the average score would have been if all your participants had been in the experimental group. Similarly, you will assume that the control group's average score is a good estimate of what the average score would have been if all your participants had been in the control group. Therefore, the first step in analyzing your data will be to calculate the average score for each group. Usually, the average you will calculate is the **mean**: the result of adding up all the scores and then dividing by the number of scores (e.g., the mean of 3 and 5 is 4 because $3 + 5 = 8$ and $8/2 = 4$).

Comparing Sample Means: How to Compare Two Imperfect Estimates

Once you have your two sample means, you can compare them. Before talking about how to compare them, let's understand why we are comparing the means. We are comparing the sample means because we know that, before we administered the treatment, both groups represented a random sample of the population consisting of every participant who took part in the study. Thus, at the end of the experiment, if the treatment had no effect, the control and experimental groups would both still be random samples from that population.

As you know, two random samples from the same population will probably be similar to each other. For instance, two random samples of the entire population of New York City should be similar to each other, two random samples from the entire population of students at your school should be similar to each other, and two random samples from the entire group of participants who took part in your study should be similar to each other. Consequently, if the treatment has no effect, at the end of the experiment, the experimental and control groups should be similar to each other.

Why We Must Do More Than Subtract the Means From Each Other

Because two random samples from the same population should be similar to each other, you might think all we need to do is subtract the control group mean from the experimental group mean to find the effect. But such is not the case: Even if the treatment has no effect, the means for the control group and experimental group will rarely be identical. To illustrate, suppose that Dr. N. Ept made a serious mistake while trying to do a double-blind study. Specifically, Dr. N. Ept succeeded in not letting his assistants know whether the participants were getting the real treatment or a placebo, but failed in that all the participants got the placebo. In other words, both groups ended up being random samples of the same population—participants who did not get the treatment. Even in such a case, the two groups will probably have different means.

How Random Error Affects Data From the Simple Experiment

Dr. N. Ept's study illustrates an important point: Even if groups are random samples of the same population, they may still differ because of random error. You are probably aware of random error from reading about public opinion polls that admit to a certain degree of sampling error.

To help you see how random error could affect the results of a simple experiment, let's simulate conducting a small-scale experiment. Be warned that this simulation won't show us what would typically happen in an experiment. Instead, this simulation is rigged to demonstrate the worst random error can do. Nevertheless, the simulation does demonstrate a fundamental truth: Random error alone can create groups that differ substantially from each other.

To conduct this simulation, assume that you have the following four participants, who would tend to score as follows:

Abby	10
John	20
Mary	70
Paul	40

Now use Box 10.1 to randomly assign each participant to either the experimental or control group. Then, get an average for each group. Repeat this process several times. If you do this, you will simulate what happens when you do an experiment and the treatment has no effect.

As doing this simulation will reveal, which participants end up in which group varies greatly depending on where on the random numbers table you happen to start—and there are many different places you could start. Not all of these possible ways of splitting participants into control and experimental groups are going to produce identical groups. Indeed, you may even find that random assignment sometimes results in having all men in the experimental group and all women in the control group.

In summary, the control and experimental groups start off as random samples of your participants. At the start of the study, these groups are not identical. Instead, they will probably merely be similar. Occasionally, however, they may start off being fairly different. If they start off as different, then they may score differently on the dependent measure task at the end of the experiment—even when the treatment has no effect. Thus, even if the treatment had no effect, random error might make the experimental group score differently (either higher or lower) than the control group.

Because random error can affect the results of a study, you need to understand random error to understand the results of a study. More specifically, to interpret the results of a simple experiment, you need to understand two important statistical principles:

1. Random error affects individual scores.
2. Random error may also cause group means to differ.

Fortunately, as you will soon see, you already intuitively understand both of these principles.

Random Error Makes Scores Within a Group Differ

To see that you intuitively grasp the first principle (random error affects individual scores), consider the following scores:

CONTROL	EXPERIMENTAL
70	80
70	80
70	80

Is there something strange about these data? Most students we show these data to realize that these data are faked. Students are suspicious of these data because scores within each group do *not* vary: There are no within-groups differences in this experiment. These data make it look like the only thing that affects scores is the treatment. With real data, however, scores would be affected by nontreatment factors. Consequently, the scores within each group would vary. That is, there would be what statisticians call within-groups variability.

When asked to be more specific about why they think the data are faked, students point out that there are at least two reasons why scores within each group should differ. First, participants within each group differ from each other, so their scores would reflect those differences. That is, because participants in the control group aren't all clones of each other, their scores won't all be the same. Likewise, because participants in the experimental group aren't all identical, their scores shouldn't all be identical.

Second, even if a group's participants were all identical, random measurement errors alone would prevent participants from getting identical scores. For instance, even if the control group participants were clones, participants' scores would probably vary due to the measure's less-than-perfect reliability. Similarly, even if all the experimental group participants were identical, their scores would not be: Many random factors—from random variations in how the experimenter treated each participant to random errors in coding of the data—would inevitably cause scores within the experimental group to differ.

In summary, most students have an intuitive understanding that there will be differences within each group (within-groups variability), and these differences are due to factors completely unrelated to the treatment. To be more specific, these differences are due to random error caused by such factors as individual differences, random measurement error, and imperfect standardization.

Random Error Can Make Group Means Differ

To see whether you intuitively grasp the second principle (random error may cause group means to differ from each other), consider the following data:

CONTROL	EXPERIMENTAL
70	70
80	80
70	100

Do you think the experimental group is scoring significantly higher than the control group? Most students wisely say "no." They realize that if the participant who scored "100" had been randomly assigned to the control group rather than the experimental group, the results may have been completely different. Thus, even though the group means differ, the difference may not be due to the treatment. Instead, the difference between these two group means could be entirely due to random error.

As you have just seen, even if the treatment has no effect, random error may cause the experimental group mean to differ from the control group mean. Therefore, we cannot say that there is a treatment effect just because there is a difference between the experimental group's average score and the control group's. Instead, if we are going to find evidence for a treatment effect, we need a difference between our groups that is "too big" to be due to random error alone.

When Is a Difference Too Big to Be Due to Random Error?

What will help us determine whether the difference between group means is too big to be due to random error alone? In other words, what will help us determine that the treatment had a statistically significant (reliable) effect?

To answer the question of how we determine whether the treatment had a statistically significant effect, we'll look at three sets of experiments. Let's begin with the two experiments tabled below. Which of the following two experiments do you think is more likely to reveal a significant treatment effect?

EXPERIMENT A		EXPERIMENT B	
Control	Experimental	Control	Experimental
70	70	70	80
71	73	71	81
72	72	72	82

Bigger Differences Are Less Likely to Be Due to Chance Alone

If you picked Experiment B, you're right! All other things being equal, bigger differences are more likely to be "too big to be due to chance alone" than smaller differences. Therefore, bigger differences are more likely to reflect a treatment effect. Smaller differences, on the other hand, provide less evidence of a treatment effect.

To appreciate the fact that small differences provide less evidence of a treatment effect, let's consider an extreme case. Specifically, let's think about the case where the difference between groups is as small as possible: zero. In that case, the control and experimental groups would have identical means. If the treatment group's mean is the same as the no-treatment group's mean, there's no evidence of a treatment effect.

"Too Big to Be Due to Chance" Partly Depends on How Big "Chance" Is

You have seen that the difference between means is one factor that affects whether a result is statistically significant. All other things being equal, bigger differences are more likely to be significant.

The size of the difference isn't the only factor that determines whether a result is too big to be due to chance. To illustrate this fact, compare the two experiments below. Then, ask yourself, is Experiment A or Experiment B more likely to reveal a significant treatment effect? That is, in which experiment is the difference more likely to be too big to be due to chance?

EXPERIMENT A		EXPERIMENT B	
Control	Experimental	Control	Experimental
68	78	70	70
70	80	80	80
72	82	60	90

Differences Within Groups Tell You How Big Chance Is

In both experiments, the difference between the experimental and control group mean is 10. Therefore, you can't tell which difference is more likely to be too big to be due to chance just by seeing which experiment has a bigger difference between group means. Instead, to make the right choice, you have to figure out the answer to this question: "In which experiment is chance alone a less likely explanation for the 10-point difference?"

To help you answer this question, we'll give you a hint. The key to answering this question correctly is to look at the extent to which scores vary within each group. The more variability within a group, the more random error is influencing scores. All other things being equal, the more random error makes individual scores within a group differ from one another (i.e., the bigger the within-groups variability), the more random error will tend to make group means differ from each other.

Now that you've had a hint, which experiment did you pick as being more likely to be significant? If you picked Experiment A, you're correct!

If you were asked why you picked A instead of B, you might say something like the following: "In Experiment B, the experimental group may be scoring higher than the control group merely because the participant who scored a 90 randomly ended up in the experimental group rather than in the control group. Consequently, in Experiment B, the difference between the groups could easily be due to random error."

Such an explanation is accurate, but too modest. Let's list the four steps of your reasoning:

1. You realized that there was more variability within each group in Experiment B than in Experiment A. That is, in Experiment B relative to Experiment A, (1) control group scores were further from the control group mean, and (2) experimental group scores were further from the experimental group mean.

2. You recognized that within-groups variability could not be due to the treatment. You realized that the differences among participants' scores within the control group could not be due to the treatment because none of those participants received the treatment. You also realized that the differences among scores within the experimental group could not be due to the treatment because every participant in the experimental group received the same treatment. Therefore, when scores within a group vary, these differences must be due to nontreatment factors such as individual differences.
3. You realized that random assignment turned the variability due to nontreatment factors (such as individual differences) into random error. Thus, you realized that the greater within-groups variability in Experiment B meant there was more random error in Experiment B than in Experiment A.
4. You realized that the same random error that caused differences within groups could cause differences between groups. That is, the more random error is spreading apart scores within each group, the more random error could be spreading the groups apart.

As you have seen, all other things being equal, the *larger* the *differences between* your *group means*, the *more likely* the results are to be *statistically significant*. As you have also seen, the s*maller* the *differences* among scores *within* each of your *groups* (i.e., the less your individual scores are influenced by random error), the *more likely* your results are to be *statistically significant*. Thus, you have learned two of the three factors that determine whether a difference is significant. To find out what the third factor is, compare Experiments A and B below. Which is more likely to produce a significant result?

EXPERIMENT A

Control	Experimental
68	70
70	72
72	74

EXPERIMENT B

Control	Experimental
68	70
70	72
72	74
68	70
70	72
72	74
68	70
70	72
72	74

In both experiments, the group means are equally far apart, so you can't look at group differences to figure out which experiment is more likely to be significant. In both experiments, the random variability within each group is the same; therefore, looking at within-groups variability will not help you figure out which experiment is more likely to be significant. Which one do you choose?

With Larger Samples, Random Error Tends to Balance Out

If you chose Experiment B, you're correct! Experiment B is the right choice because it had more participants. In Experiment B, it's less likely that random error alone would cause the groups to differ by much because *with large enough samples, random error tends to balance out to zero*. If you flip a coin 4 times, you are likely to get either 75% heads or 75% tails. That is, random error alone will probably cause a deviation of 25% or more from the true value of 50% heads. If, on the other hand, you flip a coin 4,000 times, you will almost never get more than 51% heads or fewer than 49% heads. Because 4,000 flips gives random error an opportunity to balance out, random error will almost never cause a deviation of even 1% from the true value.

Just as having more coin flips allows more opportunities for the effects of random error to balance out, having more participants allows more opportunities for random error to balance out. Thus, Experiment B, by having more participants, does a better job than Experiment A at allowing the effects of random error to balance out. Consequently, it's less likely that random error alone would cause Experiment B's groups to differ by a large amount. Therefore, a difference between the control group mean and the treatment group mean that would be big enough to be statistically significant (reliable) in Experiment B might *not* be significant in Experiment A.

ANALYZING THE RESULTS OF THE SIMPLE EXPERIMENT: THE *t* TEST

To determine whether a difference between two group means is significant, researchers often use either ANOVA[8] (analysis of variance, a technique we will discuss in the next chapter) or the ***t* test** (to see how to do a *t* test, you can use the formula in Table 10.6 or consult Appendix E).[9] Although we have not yet talked about the *t* test, you already understand the basic logic behind it. The basic idea behind the *t* test is to see *whether the difference between two groups is larger than would be expected by random error alone.* Thus, you should not be surprised to find that the *t* ratio takes the

[8] The logic of ANOVA is similar to that of the *t* test. Indeed, for a simple experiment, the *p* value for the ANOVA test will be exactly the same as the *p* value from the *t* test. Thus, if the *t* test is statistically significant (*p* is less than .05), the ANOVA test will also be statistically significant (*p* will be less than .05). In addition, for the simple experiment, you can get the value of the ANOVA test statistic (called "F") by squaring your *t* value. Thus, if *t* is 2, *F* will be 4. To learn more about ANOVA, see the next chapter or see Appendix E.

[9] Although *t* test and ANOVA analyses are commonly used, they are criticized. The problem is that both *t* tests and ANOVA tell us only whether a result is statistically significant—and, as we discussed earlier, nonsignificant results don't tell you anything and significant results don't tell you anything about the size of your effect. Therefore, many argue that, rather than using significance tests, researchers should use confidence intervals. For more on the statistical significance controversy, see Box 1 in Appendix E. For more about confidence intervals, see Appendix E.

TABLE **10.6**
Basic Idea of the *t* Test

GENERAL IDEA	FORMULA
Top of *t* ratio: Obtain observed difference (between two group means) Bottom of *t* ratio: Estimate difference expected by chance (using the standard error of the difference between means)	$t = \frac{\text{Group 1 Mean} - \text{Group 2 Mean}}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}}$ where S_1 = standard deviation of Group 1, S_2 = standard deviation of Group 2, N_1 = number of participants in Group 1, and N_2 = number of participants in Group 2. The standard deviation can be calculated by the formula $S = \sqrt{(\Sigma X - M)^2 / N - 1}$ where X stands for the individual scores, M is the sample mean, and N is the number of scores.

Notes:

1. A large *t* value is likely to be statistically significant. That is, a large *t* (above 2.6) is likely to result in a *p* value smaller than .05.
2. *t* will tend to be large when
   a. The difference between experimental group mean and the control group mean is large.
   b. The standard error of the difference is small. The standard error of the difference will tend to be small when
      i. The standard deviations of the groups are small (scores in the control group tended to stay close to the control group mean, scores in the experimental group tended to stay close to the experimental group mean).
      ii. The groups are large.

difference between the group means and divides that difference by an index of the extent to which random error might cause the groups to differ. To be more precise, ***t*** equals the difference between means divided by the standard error of the difference between means (see Table 10.6).

Making Sense of the Results of a *t* Test

Once you have obtained your *t* value, you should calculate the degrees of freedom for that *t*. To calculate degrees of freedom, subtract 2 from the number of participants. Thus, if you had 32 participants, you should have 30 degrees of freedom.

If you calculate *t* by hand, you need to compare your calculated *t* to a value in a *t* table (you could use Table 1 in Appendix F) to determine whether your *t* ratio is significant. To use the *t* table in Appendix F, you need to know how many degrees of freedom *(df)* you have. For example, if you had data from 32 participants, you would look at the *t* table in Appendix F under the row labeled "30 *df*." When comparing the *t* ratio you calculated to the value in the table, act like your *t* ratio is positive even if your

t value is actually negative (e.g., treat –3 as if it were +3). In other words, take the absolute value of your *t* ratio.

If the absolute value of your *t* ratio is *not* bigger than the number in the table, your results are *not* statistically significant at the $p < .05$ level. If, on the other hand, the absolute value of your *t* ratio is bigger than the number in the table, your results are statistically significant at the $p < .05$ level.

If you had a computer calculate *t* for you, make sure that the degrees of freedom *(df)* for *t* are two fewer than the number of participants. For example, if you thought you entered scores for 32 participants but your $df = 18$, you know there is a problem because the computer is acting as though you entered only 20 scores.

If you had a computer calculate *t* for you, it might provide you with only the *t,* the degrees of freedom, and the *p* value, as in the following case:

$$df = 8,\ t = 4,\ \text{ and }\ p < .0039$$

From the *df* of 8, you know that the *t* test was calculated based on scores from 10 participants (10 – 2 = 8). From the *p* value of less than .05, you know the results are statistically significant at the .05 level. That is, you know that if the null hypothesis were true, the chances of your obtaining differences between groups that were as big as or bigger than what you observed were less than 5 in 100.

Many computer programs will provide you with more information than the *df*, *t*, and *p* values. Some will provide you with what might seem like an overwhelming amount of information, such as the following:

1. $df = 8$, $t = 4$, and Sig. (2-tailed) = .0039
2. Mean difference = 4.00
3. 95% CI of this difference: 1.69 to 6.31
4. Group 1 mean = 11.00; Group 1 $SD = 1.58$; $SEM = 0.71$
5. Group 2 mean = 7.00; Group 2 $SD = 1.58$; $SEM = 0.71$

The first line tells you that the *t* test was calculated based on scores from 10 participants (10 – 2 = 8, the *df*) and that the results were statistically significant. The second line tells you that the Group 1 mean was 4 units bigger than the Group 2 mean. The third line tells you that you can be 95% confident that the true difference between the means is between 1.69 units and 6.31 units. (To learn more about how the confidence interval [CI] was calculated, see Box 10.2.)

The fourth line describes Group 1's data, and the fifth line describes Group 2's data. Both of those lines start by providing the group's average score (the mean) followed by a measure of how spread out the group's scores are: the standard deviation (*SD*). Be concerned if the *SD* of either group is extremely high—a high *SD* may mean that you have entered a wrong value (e.g., when entering responses from a 1-to-5 scale, you once typed a "55" instead of a "5"). Both lines end with their group's standard error of the mean (*SEM*): an indicator of how far off the group's sample mean is likely to be from the actual population mean. If either group's *SEM* is large, your experiment has little power, and you probably failed to find a significant effect.

Suppose that your experiment was powerful enough to find an effect that is statistically significant at the $p < .05$ level. In that case, because there's less than a 5% chance that the difference between your groups is solely due to

BOX 10.2 Beyond Statistical Significance: Obtaining Information About Effect Size

Your study's *t* value gives you almost everything you need to know to determine whether your results are statistically significant. However, you may also want to know whether your results are *practically* significant. To know that, you may need to know how large your effect is.

Using *t* to Estimate the Treatment's Average Effect: Confidence Intervals

One way to estimate effect size is to take advantage of information you used when you computed your *t*. Let's start by looking at the top of the *t* ratio: the difference between the mean of the no-treatment group and the mean of the treatment group. The top of the *t* ratio is an estimate of the treatment effect. Thus, if the treatment group scores 2 points higher than the no-treatment group, our best estimate is that the treatment improved scores by 2 points.

Unfortunately, our best estimate is almost certainly wrong: We have almost no confidence that the treatment effect is exactly 2.000. We would be more confident of being right if we said that the treatment effect was somewhere between 1 and 3 points. We would be even more confident of being right if we said that the real effect was somewhere between 0 and 4 points. What we would like to do is be more specific. We would like to say how confident we are that the real effect is within a certain range. For example, we would like to be able to say that we are "95% confident that the effect of the treatment is between 1 and 3 points."

Fortunately, we can specify that we are 95% confident that the real effect is between two values by using the information we used to execute the *t* test: the mean difference (the top of our *t* ratio), the standard error of the difference (the bottom of our *t* ratio), and the critical value of *t* at the .05 level. You can find the critical value by looking in the *t* table (Table 1 of Appendix F) at the intersection of the ".05" column and the row corresponding to your experiment's degrees of freedom. For example, if you had data from 42 participants, the value would be 2.021.

The middle of our confidence interval will be the difference between the means of the treatment group and the no-treatment group. That is, it will be the top of the *t* ratio. In this example, that difference is 2. To get our confidence interval's upper value, we start with the difference between our means (2). Then, we add the number we get by multiplying the standard error of the difference (the bottom of our *t* value) by the critical value of *t*. To illustrate, suppose that the difference between our means was 2, the standard error of the difference was 1, and the critical value of *t* was 2.021. To 2, we would add 2.021 (the standard error of the difference [1] × the critical value of *t* [2.021] = 2.021). Thus, the upper value of our confidence interval would be 4.021 (2 + 2.021).

To get the lower value, we reverse the process. We will again start with 2 (the difference between our means). This time, however, we will subtract, rather than add, 2.021 (the product of multiplying the standard error by the critical *t* value) from 2. Therefore, the lower value of our interval would be −0.021 [2 − (1 × 2.021) = 2 − 2.021 = −0.021].

As the result of our calculations, we could say that we were 95% confident that the true effect was in the interval ranging from −0.021 to 4.021. By examining this interval, we can form two conclusions. First, we cannot confidently say that the treatment effect has any effect because 0 (zero effect, no effect) was within our interval. Second, we see that our confidence interval is large and so our study lacks power and precision. Therefore, we may want to repeat the study in a way that shrinks the confidence interval (e.g., using more participants, using more reliable measures, using more homogeneous participants, using more standardized procedures) so that we can more precisely estimate the treatment's effect.

For example, in the original study, we studied 42 participants. If we repeated the study using 62 participants and again found a difference between our groups of 2, we would be 95% confident that the true effect was between .35 and 3.6.[1] Not only is this interval narrower than the original interval

[1] When we calculated this confidence interval, we assumed that the standard deviations (an index of the extent to which participants' scores differ from the mean; a 0 would mean that nobody's score differed from the mean) within each of your groups would be the same as they were in the original study. If your procedures were more standardized when you repeated the study, the standard deviations might be smaller and so your intervals might be even smaller than what we projected.

(Continued)

BOX 10.2 Continued

(which went from –0.021 to +4.021), but it also does not include zero. Therefore, we could confidently say that the treatment did have some effect. Note another lesson from this example: Even though the first study's results were not statistically significant (because we could not say that the treatment effect was significantly different from zero) and the second study's results were significant (because we could say that the treatment effect was significantly different from zero), the two studies do not contradict each other. The difference in the results is that the second study, by virtue of its greater power and precision, allows you to make a better case that the treatment effect is greater than zero.

Using *t* to Compute Other Measures of Effect Size: Cohen's *d* and r^2

In the previous section, you learned how to provide a range that you were 95% confident contained the average effect of the treatment. However, even if you knew precisely what the average effect of the treatment was, you would not know all you should know about the treatment's effect size. For example, suppose you know that the average effect was 2. Is 2 a small effect? If your participants' scores range from 0 to 100, a difference between your control group and experimental group of 2 units might be a relatively small effect. If, on the other hand, scores in your control group vary from 0 to 1, and scores in your treatment group vary from 2 to 3, a treatment effect of 2 units would be a relatively large effect. Therefore, to know the relative size of an effect, you need an effect size measure that takes into account the variability of the scores.

One popular effect size measure is **Cohen's *d*.** If you had the same number of participants in each group, you can calculate Cohen's *d* from your *t* value by using the following formula: Cohen's $d = 2t/\sqrt{df}$. Thus, if *t* is 3 and *df* is 9, Cohen's *d* will be $(2\times3)/\sqrt{9} = 6/3 = 2$. Usually, social scientists view a *d* of 0.2 as indicating a small effect, a *d* of 0.5 as indicating a medium effect, and a *d* of 0.8 as indicating a large effect.

Another way of measuring the relationship between the treatment and your dependent variable is to square the correlation (r) between the treatment and the dependent variable. The result will be a measure, called the **coefficient of determination**, that can range from 0 (no relationship) to 1.00 (perfect relationship). Usually, social scientists view a coefficient of determination of .01 as small, of .09 as moderate, and of .25 as large (for more about the coefficient of determination, see Chapter 7). If you have computed *d,* you can compute the coefficient of determination (r^2) by using the following formula: $r^2 = d^2/(d^2 + 4)$. To see the relationships among these effect size measures, see Table 10.7.

chance, you can be reasonably sure that some of the difference is due to your treatment.

To learn about the size of your treatment's effect, you might want to use Box 10.2 to compute an index of effect size such as Cohen's *d*. For example, suppose your computer analysis presented the following results:

1. $df = 30$, $t = 3.10$, and $p < .05$
2. Mean difference = 3.46
3. 95% CI of this difference: 1.57 to 5.35; $SED = 1.12$
4. Group 1 mean = 8.12; Group 1 $SD = 3.0$; $SEM = 0.75$
5. Group 2 mean = 4.66; Group 2 $SD = 3.32$; $SEM = 0.83$

TABLE **10.7**
Relationship Among Different Effect Size Measures

INFORMATION FROM THE *T* TEST			EFFECT SIZE MEASURES		
t	Degrees of Freedom	Mean Difference (example with low variability in scores)	Mean Difference (example with moderate variability in scores)	*d*	r^2 (also called h^2)
2	9	2	4.7	1.33	.31
2	16	1.4	3.7	1.0	.20
2	25	1.2	3.0	0.8	.14
2	36	1.0	2.5	0.67	.10
2	49	0.8	2.2	0.57	.08
2	64	0.7	1.9	0.50	.06
2	81	0.7	1.7	0.44	.05
2	100	0.6	1.5	0.40	.04

Using that data and Box 10.2, you would be able to determine that Cohen's *d* was 1.13.

Then, you could write up your results as follows:[10] "As predicted, the experimental group recalled significantly more words ($M = 8.12$, $SD = 3.0$) than the control group ($M = 4.66$, $SD = 3.32$), $t(30) = 3.10$, $p < .05$, $d = 1.13$."

You could include even more information: APA strongly encourages researchers to supplement significance tests with means, standard deviations, and both confidence intervals and effect size measures. However, at the very least, you should say something like this: "As predicted, the experimental group recalled significantly more words ($M = 8.12$) than the control group ($M = 4.66$), $t(30) = 3.10$, $p < .05$."

You must do more than report that your results are statistically significant. Indeed, largely because some researchers have focused only on whether their results are statistically significant, a few researchers have suggested that statistical significance testing be banned (for more on the statistical significance controversy, see Box 1 in Appendix E). Although not everyone agrees that statistical significance testing should be banned, almost everyone agrees that researchers need to do more than report *p* values.

[10] *M* stands for mean, *SD* stands for standard deviation (a measure of the variability of the scores; the bigger the SD, the more spread out the scores are and the less the scores cluster around the mean), and *d* stands for Cohen's *d* (a measure of effect size). *SD* will usually be calculated as part of computing *t* (for more about *SD*, see Appendix E). To learn how to compute *d*, see Box 10.2.

Assumptions of the *t* Test

The validity of any *p* values you obtain from any significance test will depend on how well you meet the assumptions of that statistical test. For the *t* test, two of these assumptions are especially important: (1) having at least interval scale data and (2) having independent observations.

Two Critical Assumptions

When the *t* test determines whether one group's mean score is significantly larger than the other's, it assumes that groups with higher means have more of the quality you are measuring than groups with lower means. Because only interval and ratio scale data allow you to compute such "meaningful means," you must be able to assume that you have either interval scale or ratio scale data (for a review of interval and ratio scale data, see Chapter 6).

Because you cannot compute meaningful means on either qualitative data or ranked data, you cannot do a *t* test on those data. You cannot compute meaningful means on qualitative (nominal, categorical) data because scores relate to categories rather than amounts. With qualitative (nominal) data, 1 might equal "nodded head," 2 might equal "gazed intently," and 3 might equal "blinked eyes." With such nominal data, computing a mean (e.g., the mean response was 1.8) would be meaningless.

With ranked and other ordinal data, the numbers have an order, but they still don't refer to specific amounts and so means can be meaningless and misleading. For example, although averaging the ranks of second- and third-place finishers in a race would result in the same mean rank (2.5) as averaging the ranks of the first- and fourth-place finishers, the mean times of the two groups might be very different. Despite having the same average rank, the average times of the first- and fourth-place finishers could be much faster or much slower than the average of the times of the second- and third-place finishers.

Although having either nominal or ordinal data prevents you from comparing group means with a *t* test, you can still compare two groups using tests, such as the Mann-Whitney U test (for ordinal data) and the chi-square test (for either nominal or ordinal data), that do not involve comparing means. (For more on these tests, see Appendix E.)

The second assumption you must meet to perform a legitimate *t* test is that your observations must be independent. Specifically, (a) participants must be assigned independently (e.g., individually, so that the assignment of Mary to the experimental group has no effect on whether John is assigned to the experimental group); (b) participants must respond independently (e.g., no participant's response influences any other participant's response); and (c) participants must be tested independently so that, other than the treatment, there is no systematic difference between how experimental and control group participants are treated.

If you followed our advice and independently and randomly assigned each participant to either the experimental or the control conditions, and then ran participants individually (or in small groups or in larger groups that mixed experimental and control participants), your observations are independent. If, however, your observations are not independent, you cannot legitimately do a conventional independent groups *t* test. Indeed, violating

TABLE **10.8**
Effects of Violating the *t* Test's Assumptions

ASSUMPTION	CONSEQUENCES OF VIOLATING ASSUMPTION
Observations are independent (participants are independently assigned and participants do not influence one another's responses).	Serious violation; probably nothing can be done to salvage your study.
Data are interval or ratio scale (e.g., numbers must not represent qualitative categories, nor may they represent ranks [first, second, third, etc.]).	Do not use a *t* test. However, you may be able to use another statistical test (e.g., Mann-Whitney U, Chi-square).
The population from which your sample means was drawn is normally distributed.	If the study used more than 30 participants per group, this is not a serious problem. If, however, fewer participants were used, you may decide to use a different statistical test.
Scores in both conditions have the same variance.	Usually not a serious problem.

independence often means that the data from your study are unanalyzable and thus worthless.

To reiterate, to do a meaningful independent *t* test in a simple experiment, your data must meet two key assumptions: You must have at least interval scale data, and you must have used independently assigned participants to groups. In addition to these two pivotal assumptions, the *t* test makes two less vital assumptions (see Table 10.8).

Two Less Critical Assumptions

First, the *t* test assumes that the individual scores in the population from which your sample means were drawn are **normally distributed**: half the scores are below the average score; half are above; the average score is the most common score; about 2/3 of the scores are within one standard deviation of the mean; about 19/20 of the scores are within two standard deviations of the mean; and if you were to plot how often each score occurred, your plot would resemble a bell-shaped curve. The reason for this assumption is that if the individual scores in the population are normally distributed, the distribution of sample means based on those scores will also tend to be normally distributed.[11] The assumption that individual scores are normally

[11] Why do we have to assume that the distribution of sample means is normally distributed? We need to know precisely how the sample means are distributed to establish how likely it is that the two sample means could differ by as much as they did by chance alone. In other words, if we are wrong about how the sample means are distributed, our *p* value—our estimate of the probability of the sample means differing by as much as they did if their population means were the same—would be wrong.

distributed is usually nothing to worry about because most distributions are normally distributed.

But what if the individual scores aren't normally distributed? Even then, your sample means probably will be normally distributed—provided you have more than 30 participants per group. That is, as the **central limit theorem** states, with large enough samples (and 30 per group is usually large enough), the distribution of sample means will be normally distributed, regardless of how individual scores are distributed.

To understand why the central limit theorem works, realize that if you take numerous large samples from the same population, your sample means will differ from one another for only one reason: random error. Because random error is normally distributed, the distributions of sample means will be normally distributed—regardless of the shape of the underlying population.

The *t* test's second less critical assumption is that the variability of scores within your experimental group will be about the same as the variability of scores within your control group. To be more precise, the assumption is that scores in both conditions will have the same variance.[12] Usually, the penalty for violating the assumption of equal variances is not severe. Specifically, if you have unequal variances, it won't seriously affect the results of your *t* test, as long as one variance isn't more than 2½ times larger than the other.

QUESTIONS RAISED BY RESULTS

Obviously, if you violate key assumptions of the *t* test, people should question your results. But even if you don't violate any of the *t* test's assumptions, your results will raise questions—and this is true whether or not your results are statistically significant.

Questions Raised by Nonsignificant Results

Nonsignificant results raise questions because the null hypothesis cannot be proven. Therefore, null results inspire questions about the experiment's power such as the following:

1. Did you have enough participants?
2. Were the participants homogeneous enough?
3. Was the experiment sufficiently standardized?
4. Were the data coded carefully?
5. Was the dependent variable sensitive and reliable enough?
6. Would you have found an effect if you had chosen two different levels of the independent variable?

---

[12] To get the variance for a group, square that group's standard deviation (*SD*). If you used a computer to get your *t*, the computer program probably displayed each group's *SD*. If you calculated the *t* by hand, you probably calculated each group's *SD* as part of those calculations. Some computer programs will do a statistical test such as Levene's Test for Equality of Variance to tell you how reasonable it is to assume that the groups have the same variance. If the *p* value for the Levene's Test for Equality of Variance is statistically significant, it means that the variances are probably different: It does *not* mean that the treatment has an effect. If the variances are significantly different, instead of a conventional *t* test, you may want to do Welch's test instead. Some programs will also calculate two *t* values for you: one assuming equal variances, one not making that assumption.

Questions Raised by Significant Results

If your results are statistically significant, it means you found an effect for your treatment. So, there's no need to question your study's power. However, a significant effect raises other questions. Sometimes, questions are raised because statistical significance doesn't tell us how big the effect is (see Box 10.2).

Sometimes, questions are raised because the experimenter sacrificed construct or external validity to obtain adequate power. For example, if you used an empty control group, you have questionable construct validity. Consequently, one question would be: "Does your significant treatment effect represent an effect for the construct you tried to manipulate or would a placebo treatment have had the same effect?" Or, if you used an extremely homogeneous group of participants, the external validity of your study might be questioned. For instance, skeptics might ask: "Do your results apply to other kinds of participants?" Thus, skeptics might want you to increase the external validity of your study by repeating it with a more representative sample. Specifically, they might want you to first use random sampling to obtain a representative group of participants and then randomly assign those participants to either the control or experimental group.

At other times, questions are raised because of a serious limitation of the simple experiment: It can study only two levels of a single independent variable. Because of this, there are two important questions you can ask of any simple experiment:

1. To what extent do the results apply to levels of the independent variable that were not tested?
2. To what extent could the presence of other variables modify (strengthen, weaken, or reverse) the treatment's effect?

CONCLUDING REMARKS

As you have seen, the results of a simple experiment always raise questions. Although results from any research study raise questions, some questions raised by the results of the simple experiment occur because the simple experiment is limited to studying only two levels of a single variable. If the logic of the simple experiment could be used to create designs that would study several levels of several independent variables, such designs could answer several questions at once. Fortunately, as you will see in Chapters 11 and 12, the logic of the simple experiment can be extended to produce experimental designs that will allow you to answer several research questions with a single experiment.

SUMMARY

1. Psychologists want to know the causes of behavior so that they can understand people and help people change. Only experimental methods allow us to isolate the causes of an effect.
2. Studies that don't manipulate a treatment are not experiments.
3. Many variables, such as participant's age, participant's gender, and participant's personality, can't be manipulated. Therefore,

many variables can't be studied using an experiment.

4. The simple experiment is the easiest way to establish that a treatment causes an effect.
5. The experimental hypothesis states that the treatment will cause an effect.
6. The null hypothesis, on the other hand, states that the treatment will not cause an observable effect.
7. With the null hypothesis, you only have two options: You can reject it, or you can fail to reject it. You can never accept the null hypothesis.
8. Typically, in the simple experiment, you administer a low level of the independent (treatment) variable to some of your participants (the comparison or control group) and a higher level of the independent variable to the rest of your participants (the experimental group). Near the end of the experimental session, you observe how each participant scores on the dependent variable: a measure of the participant's behavior.
9. To establish causality with a simple experiment, participants' responses must be independent. Because of the need for independence, your experimental and control groups are not really groups. Instead, these "groups" are sets of individuals.
10. Independent random assignment is the cornerstone of the simple experiment: Without it, you do not have a simple experiment.
11. Independent random assignment is necessary because it is the only way to make sure that the only differences between your groups are either due to chance or to the treatment.
12. Independent random assignment makes it likely that your control group is a fair comparison group. Therefore, if you use random assignment, the control and experimental groups should be equivalent before you introduce the treatment.
13. Random assignment can be used only if you are manipulating (assigning) a treatment. It involves assigning one level of a treatment to some participants and a different level of that treatment to other participants. Random assignment helps a study's internal validity.
14. Your goal in using independent random assignment is to create two samples that accurately represent your entire population of participants. You use the mean of the control group as an estimate of what would have happened if all your participants had been in the control group. You use the experimental group mean as an estimate of what the mean would have been if all your participants had been in the experimental group.
15. The *t* test tries to answer the question, "Does the treatment have an effect?" In other words, would participants have scored differently had they all been in the experimental group than if they had all been in the control group?
16. If the results of the *t* test are statistically significant, the difference between your groups is greater than would be expected by chance (random error) alone. Therefore, you reject the null hypothesis and conclude that your treatment has an effect. Note, however, that statistical significance does not tell you that your results are big, important, or of any practical significance.
17. There are two kinds of errors you might make when attempting to decide whether a result is statistically significant. Type 1 errors occur when you mistake a chance difference for a treatment effect. Before the study starts, you choose your "false alarm" risk (risk of making a Type 1 error). Most researchers decide to take a 5% risk. Type 2 errors occur when you fail to realize that the difference between your groups is not solely due to chance. In a sense, Type 2 errors involve overlooking a genuine treatment effect.
18. By reducing your risk of making a Type 1 error, you increase your risk of making a Type 2 error. That is, by reducing your chances of falsely "crying wolf" when there is no treatment effect, you increase your chances of failing to yell "wolf" when there really is a treatment effect.
19. Because Type 2 errors can easily occur, nonsignificant results are inconclusive results.

20. To prevent Type 2 errors, (a) reduce random error, (b) use many participants to balance out the effects of random error, and (c) try to increase the size of your treatment effect.
21. You can easily determine your risks of a Type 1 error, but there's no way you can design your experiment to reduce them. In contrast, it is hard to determine your risk of making a Type 2 error, but there are many ways you can design your study to reduce your risk of making such errors.
22. If your experiment minimizes the risk of making Type 2 errors, your experiment has power. In the simple experiment, *power* refers to the ability to obtain statistically significant results when your independent variable really does have an effect.
23. Sometimes, efforts to improve power may hurt the study's external validity. For example, to get power, researchers may use a highly controlled lab setting rather than a real-life setting. Likewise, power-hungry researchers may study participants who are very similar to each other rather than a wide range of participants.
24. Occasionally, efforts to improve power may hurt the study's construct validity.
25. Using placebo treatments, single blinds, and double blinds can improve your study's construct validity.
26. Ethical concerns may temper your search for power—or even cause you to decide not to conduct your experiment.
27. Because of random error, you cannot determine whether your treatment had an effect simply by subtracting your experimental group mean from your control group mean. Instead, you must determine whether the difference between your group means could be due to random error.
28. The *t* test involves dividing the difference between means by an estimate of the degree to which the groups would differ when the treatment had no effect. More specifically, the formula for the *t* test is: (Mean 1 – Mean 2)/standard error of the difference.
29. The degrees of freedom for a two-group between-subjects *t* test are 2 less than the total number of participants.
30. The *t* test is a common way to analyze data from a simple experiment.
31. If your data do not meet the assumptions of the *t* test, your statistical analysis may give you misleading results.

KEY TERMS

internal validity (*p. 335*)
simple experiment (*p. 335*)
independent random assignment (*p. 336*)
experimental hypothesis (*p. 337*)
null hypothesis (*p. 337*)
independent variable (*p. 341*)
levels of an independent variable (*p. 341*)
experimental group (*p. 341*)
control group (*p. 341*)
independently, independence (*p. 342*)
dependent variable (dependent measure) (*p. 345*)
inferential statistics (*p. 346*)
statistical significance (*p. 346*)
null results (nonsignificant results) (*p. 347*)
Type 1 error (*p. 350*)
Type 2 error (*p. 352*)
power (*p. 352*)
empty control group (*p. 358*)
placebo treatment (*p. 358*)
blind (masked) (*p. 358*)
single blinds (*p. 358*)
double blinds (*p. 358*)
populations (*p. 360*)
mean (*p. 362*)
t test (*p. 368*)
$p < .05$ level (*p. 370*)
Cohen's *d* (*p. 372*)
coefficient of determination (*p. 372*)
normally distributed, normal distribution (*p. 375*)
central limit theorem (*p. 376*)

EXERCISES

1. A professor has a class of 40 students. Half of the students chose to take a test after every chapter (chapter test condition) outside of class. The other half of the students chose to take in-class "unit tests." Unit tests covered four chapters. The professor finds no statistically significant differences between the groups on their scores on a comprehensive final exam. The professor then concludes that type of testing does not affect performance.
   a. Is this an experiment?
   b. Is the professor's conclusion reasonable? Why or why not?
2. Participants are randomly assigned to meditation or no-meditation condition. The meditation group meditates three times a week. The meditation group reports being significantly more energetic than the no-meditation group.
   a. Why might the results of this experiment be less clear-cut than they appear?
   b. How would you improve this experiment?
3. Theresa fails to find a significant difference between her control group and her experimental group $t\ (10) = 2.11$, not significant.
   a. Given that her results are not significant, what—if anything—would you advise her to conclude?
   b. What would you advise her to do? (Hint: You know that her t test, based on 10 degrees of freedom, was not significant. What does the fact that she has 10 degrees of freedom tell you about her study's sample size, and what does it suggest about her study's power?)
4. A training program significantly improves worker performance. What should you know before advising a company to invest in such a training program?
5. Jerry's control group is the football team, his experimental group is the baseball team. He assigned the groups to condition using random assignment. Is there a problem with Jerry's experiment? If so, what is it? Why is it a problem?
6. Students were randomly assigned to two different strategies of studying for an exam. One group used visual imagery, the other group was told to study the normal way. The visual imagery group scores 88% on the test as compared to 76% for the control group. This difference was not significant.
   a. What, if anything, can the experimenter conclude?
   b. If the difference had been significant, what would you have concluded?
   c. "To be sure that they are studying the way they should, why don't you have the imagery people form one study group and have the control group form another study group?" Is this good advice? Why or why not?
   d. "Just get a sample of students who typically use imagery and compare them to a sample of students who don't use imagery. That will do the same thing as random assignment." Is this good advice? Why or why not?
7. Bob and Judy are doing the same study, except that Bob has decided to put his risk of a Type 1 error at .05 whereas Judy has put her risk of a Type 1 error at .01. (Note that consulting Table 1 in Appendix F will help you answer parts a and b.)
   a. If Judy has 22 participants in her study, what t value would she need to get significant results?
   b. If Bob has 22 participants in his study, what t value would he need to get significant results?
   c. Who is more likely to make a Type 1 error? Why?
   d. Who is more likely to make a Type 2 error? Why?
8. Gerald's dependent measure is the order in which people turned in their exam (first, second, third, etc.). Can Gerald use a t test on his data? Why or why not? What would you advise Gerald to do in future studies?
9. Are the results of Experiment A or Experiment B more likely to be significant? Why?

EXPERIMENT A		EXPERIMENT B	
Control group	Experimental group	Control group	Experimental group
3	4	0	0
4	5	4	5
5	6	8	10

10. Are the results of Experiment A or Experiment B more likely to be significant? Why?

EXPERIMENT A		EXPERIMENT B	
Control group	Experimental group	Control group	Experimental group
3	4	3	4
4	5	4	5
5	6	5	6
		3	4
		4	5
		5	6
		3	4
		4	5
		5	6

WEB RESOURCES

1. Go to the Chapter 10 section of the book's student website and
   a. Look over the concept map of the key terms.
   b. Test yourself on the key terms.
   c. Take the Chapter 10 Practice Quiz.
   d. Do the interactive end-of-chapter exercises.
2. Do a *t* test using a statistical calculator by going to the "Statistical Calculator" link.
3. Find out how to conduct a field experiment by reading "Web Appendix: Field Experiments."
4. If you want to write your method section, use the "Tips on Writing a Method Section" link.
5. If you want to write up the results of a simple experiment, click on the "Tips for Writing Results" link.

CHAPTER 11

Expanding the Simple Experiment

The Multiple-Group Experiment

Perhaps too much of everything is as bad as too little.
—Edna Ferber

Scientific principles and laws do not lie on the surface of nature. They are hidden, and must be wrested from nature by an active and elaborate technique of inquiry.
—John Dewey

CHAPTER OVERVIEW

We devoted Chapter 10 to the simple experiment: the design that involves randomly assigning participants to two groups. The simple experiment is internally valid and easy to conduct. However, it is limited in that you can study only two values of a single independent variable.

In this chapter, you will see why you might want to go beyond studying two values of a single variable. Then, you will see how the principle that gives the simple experiment internal validity (random assignment of participants to two groups) can be extended to experiments that study the effects of three or more values of a single independent variable. Finally, you will learn how to analyze data from such multiple-group experiments.

THE ADVANTAGES OF USING MORE THAN TWO VALUES OF AN INDEPENDENT VARIABLE

The simple experiment is ideal if an investigator wants to compare a single treatment group to a single no-treatment control group. However, as you will see, investigators often want to do more than compare two groups.

Comparing More Than Two Kinds of Treatments

We do not live in a world where there are only two flavors of ice cream, only two types of music, and only two opinions on how to solve any particular problem. Because people often choose between more than two options, investigators often compare more than two different kinds of treatments.

For instance, to decide how police should respond to a domestic dispute, investigators compared three different strategies: (1) arrest a member of the couple, (2) send one member away for a cooling off period, and (3) give advice and mediate the dispute (Sherman & Berk, 1984). Clearly, investigators could not compare three different treatments in one simple, two-group experiment. Therefore, instead of randomly assigning participants to two different groups, they randomly assigned participants to three different groups. (To learn how to randomly assign participants to more than two groups, see Box 11.1.)

In another case of attacking an applied problem, Cialdini (2005) saw a problem we all see—a well-intentioned, written request to do something

BOX 11.1 Randomly Assigning Participants to More Than Two Groups

Step 1 Across the top of a piece of paper write down your conditions. Under each condition draw a line for each participant you will need.

GROUP 1	GROUP 2	GROUP 3
____	____	____
____	____	____
____	____	____
____	____	____

Step 2 Turn to a random numbers table (there's one in Table 6, Appendix F). Roll a die to determine which column in the table you will use.

Step 3 Assign the first number in the column to the first space under Group 1, the second number to the second space, and so forth. When you have filled the spaces for Group 1, put the next number under the first space under Group 2. Similarly, when you fill all the spaces under Group 2, place the next number in the first space under Group 3.

GROUP 1	GROUP 2	GROUP 3
12	20	63
39	2	64
53	37	95
29	1	18

Step 4 Replace the lowest random number with "Participant 1," the second lowest random number with "Participant 2," and so on. Thus, in this example, your first two participants would be in Group 2, and your third participant would be in Group 1.

good—and he wondered what most of us have wondered: Would wording the request differently make it more effective? Specifically, he questioned the effectiveness of hotel room signs that urge guests to conserve water by reusing towels because doing so will (1) preserve the environment and (2) help the hotel donate money to an environmental cause. Cialdini believed that approaches that used psychological principles would be more effective than the hotels' usual approach, and he could think of at least two principles that he could apply.

First, he could apply the principle that people tend to do what they believe others do. Thus, he created a sign stating that 75% of guests reuse their towels.

Second, he could apply the principle that people tend to repay a favor. Thus, he created a sign stating that the hotel had already donated money to protect the environment on behalf of the hotel guests and wanted to recover that expense.

To test his two solutions against conventional practice, Cialdini needed at least three groups: (1) a group that got the conventional treatment—a "preserve the environment plus hotel donation" group, (2) a "most other people are doing it" group, and (3) a "repay a favor" group. As Cialdini suspected, both the "repay a favor" and the "most other people are doing it" reused their towels much more than the group that saw the sign hotels typically used.

Clearly, Cialdini could not compare three groups in a single, two-group experiment. Thus, he used a multiple-group experiment. Similarly, Nairne, Thompson, and Pandeirada (2007) hypothesized that people are best able to remember information when they rate its relevance to their survival. To see whether the survival rating task was the best rating task for helping participants recall information, Nairne et al. used a multiple-group experiment to compare their rating task to other rating tasks that help memory (e.g., rating how pleasurable the word is, rating how personally relevant the word is). In short, if, like Cialdini or Nairne and his colleagues, you want to compare more than two treatments, you should use a multiple-group experiment.

Comparing Two Kinds of Treatments With No Treatment

Even when you are interested in comparing only two types of treatments, you may be better off using a multiple-group experiment. To understand why, let's consider the following research finding: For certain kinds of back problems, people going to a chiropractor end up better off than those going for back surgery. Although an interesting finding, it leaves many questions unanswered. For example, is either treatment better than nothing? We don't know because the researchers didn't compare either treatment to a no-treatment control condition. It could be that both treatments are worse than nothing and chiropractic treatment is merely the lesser of two evils. On the other hand, both treatments could be substantially better than no treatment and chiropractic could be the greater of two goods.

Similarly, if we compared two untested psychological treatments in a simple experiment, we would know only which is better than the other: We would not know whether the better one was the less harmful of two "bad" treatments or the more effective of two "good" treatments. Thus, we would not know whether the lesser of the two treatments was (1) moderately harmful, (2) neither harmful nor helpful, or (3) mildly helpful. However, by using a three-group experiment that has a no-treatment control group, we would be able to judge not only how effective the two treatments were relative to each other but also their overall, general effectiveness. Consider the following examples of how adding a no-treatment control group helps us know what effect the treatments had.

- In a classic experiment, Loftus (1975) found that leading questions distorted participants' memories of a filmed car accident. All participants watched a film of a car accident, completed a test booklet that contained questions about the film, and a week later, answered some more questions about the film. But participants were not treated identically because not all participants got the same test booklet. Instead, each participant was randomly assigned to receive one of the following three test booklets:

    1. The "presume" booklet contained 40 questions asked of all participants, plus 5 additional questions that asked whether certain objects—objects that were *not* seen in the film—were in the film. These 5 additional questions were leading questions: questions suggesting that the object was shown in the film (e.g., "Did you see **the** school bus in the film?").
    2. The "mention but don't presume" booklet contained 40 questions asked of all participants, plus 5 additional questions that asked

whether certain objects—objects that were *not* seen in the film—were in the film. This booklet was the same as the "presume" booklet except that the 5 additional questions did *not* suggest that the item was shown in the film (e.g., "Did you see **a** school bus in the film?").

3. A control booklet that contained 40 questions asked of all participants.

Note that without a control group, Loftus would not have known whether the difference between the nonleading question and leading question group was due to (a) the nonleading question condition sharpening memory or (b) the leading question condition distorting memory.

- Crusco and Wetzel (1984) looked at the effects of having servers touch restaurant customers on the tips that servers received. Had they only compared hand-touching with shoulder-touching, they would not have known whether touching had an effect. Thanks to the no-touch control group, they learned that both kinds of touching increase tipping.
- Anderson, Carnagey, and Eubanks (2003) looked at the effects of violent lyrics on aggressive thoughts. Had they used only nonviolent and violent songs, they would not have known whether nonviolent songs reduced aggressive thoughts or whether violent songs increased aggressive thoughts. Thanks to the no-song control condition, they learned that violent lyrics increased aggressive thoughts.
- Strayer and Drews (2008) looked at the effects of cell phones on driving. Had they only compared the driving performance of drivers who use hand-held cell phones to drivers who use hands-free cell phones, they would not have found an effect for cell phones. However, thanks to a no cell phone control group, they learned that cell phone use impairs driving.

Comparing More Than Two Amounts of an Independent Variable to Increase External Validity

In the simple experiment, you are limited to two amounts of your independent variable. However, we do not live in a world where variables come in only two amounts. If we did, other people would be either friendly or unfriendly, attractive or unattractive, like us or unlike us, and we would be either rewarded or punished, included or excluded, and in complete control or have no control. Instead, we live in a world where situations vary not so much in terms of whether a quality (e.g., noise) is present but rather the degree to which that quality is present.

Not only that, but we live in a world where more is not always better. Sometimes, too little of some factor can be bad, too much can be bad, but (to paraphrase the littlest of the three bears) a medium amount is just right. In such cases, a simple, two-valued experiment can lead us astray.

To see how simple experiments can be misleading, suppose that a low amount of exercise leads to a poor mood, a moderate amount of exercise leads to a good mood, and a high amount of exercise leads to a poor mood. Such an upside-down "U"-shaped relationship is plotted in Figure 11.1a. As you can see, if we did a multiple-group experiment, we would uncover the true relationship between exercise and mood. However, if we did a simple

experiment, our findings might be misleading. For example, if we did the simple experiment depicted in

- Figure 11.1b, we might conclude that exercise *increases* mood
- Figure 11.1c, we might conclude that exercise *decreases* mood
- Figure 11.1d, we might conclude that exercise does not affect mood

As you have just seen, if a researcher is to make accurate statements about the effects of an independent variable, the researcher must know the independent and dependent variables' **functional relationship**: the shape of the relationship.

If you are going to map the shape of a functional relationship accurately, you need more than the two data points that a simple experiment provides.

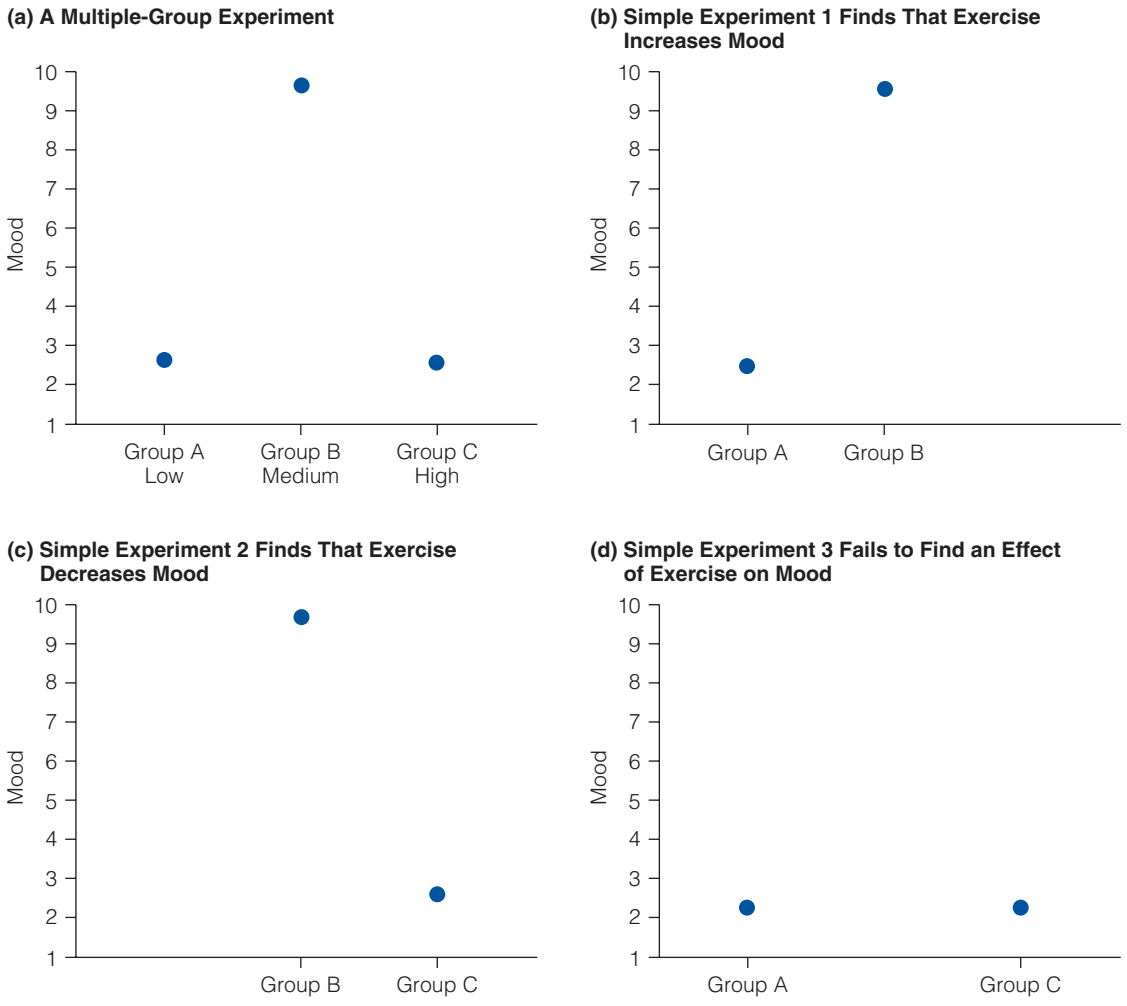

FIGURE **11.1** How a Multiple-Group Experiment Can Give You a More Accurate Picture of a Relationship Than a Simple Experiment

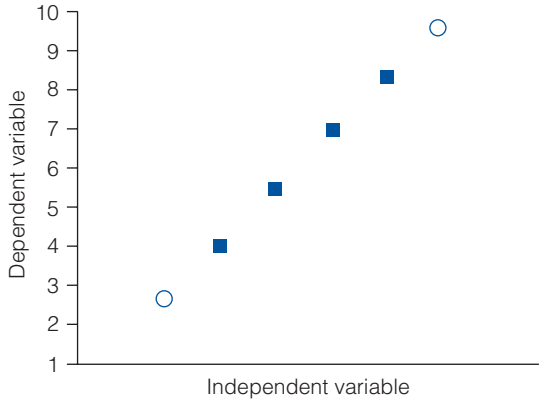

FIGURE **11.2** Linear Relationship Between Two Points

Simple experiments do not enable you to uncover the nature of a functional relationship because many different shaped lines can be drawn between two points. To appreciate this, consider Figure 11.2. From the two known data points (the empty circles), can you say what the relationship between the variables is?

No, you can't. Perhaps the relationship is a **linear relationship**: one that is represented by a straight line. A straight line does fit your two points. However, maybe your relationship is *not* linear: As you can see from Figure 11.3, many other curved lines also fit your two points.

Because lines of many different shapes can be drawn between the two points representing a simple experiment's two group means, the simple experiment does not help you discover the functional relationship between the variables. Thus, if your simple experiment indicated that 100 minutes of exercise produced a better mood than 0 minutes of exercise, you would still be clueless about the functional relationship between exercise and mood. Therefore, if we asked you about the effects of 70 minutes of exercise on mood, you could do little more than guess. If you assumed that the exercise-mood relationship is linear, you would guess that exercising 70 minutes a day would be (a) better than no exercise and (b) worse than exercising 100 minutes a day. But if your assumption of a linear relationship is wrong (and it well could be), your guess would be wrong.

To get a line on the functional relationship between variables, you need to know more than two points. Therefore, suppose you expanded the simple experiment into a multilevel experiment by adding a group that gets 50 minutes of exercise a day. Then, you would have a much clearer idea of the functional relationship between exercise and happiness. As you can see in Figure 11.4 on page 390, using three levels can help you identify the functional relationship among variables. If the relationship is linear, you should be able to draw a straight line through your three points. If the relationship is U-shaped, you'll be able to draw a "U" through your three points.

Because you can get a good picture of the functional relationship when you use three levels of the independent variable, you can make accurate

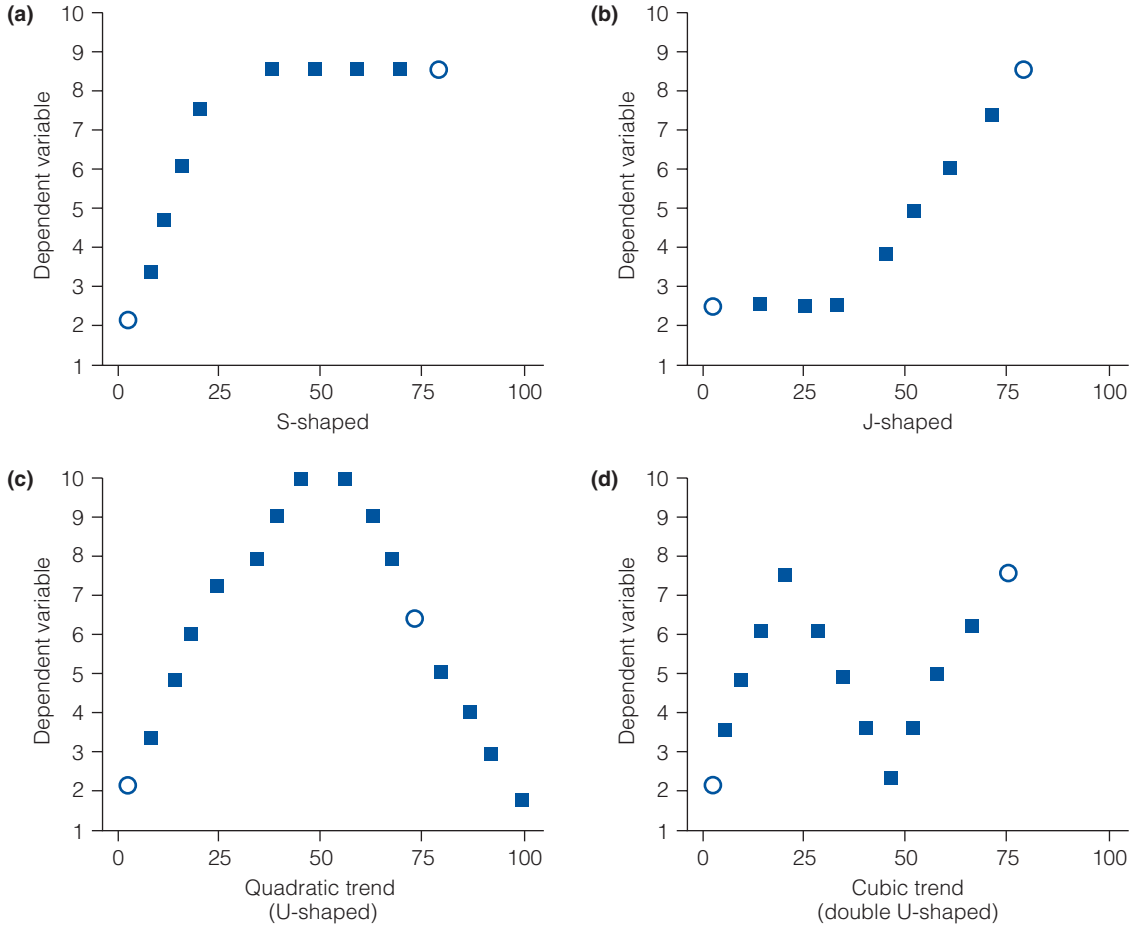

FIGURE **11.3** Some Possible Nonlinear Relationships

Note: The circles represent the known data points. The boxes between the circles are what might happen at a given level of the independent variable, depending on whether the relationship between the variables is characterized by (a) an S-shaped (negatively accelerated) trend, (b) a J-shaped (positively accelerated) trend, (c) a U-shaped (quadratic) trend, or (d) a double U-shaped (cubic) trend.

predictions about unexplored levels of the independent variable. For example, if the functional relationship between exercise and happiness were linear, you might obtain the following pattern of results:

Group 1	0 minutes of exercise per day	0.0 self-rating of happiness
Group 2	50 minutes of exercise per day	5.0 self-rating of happiness
Group 3	100 minutes of exercise per day	10.0 self-rating of happiness

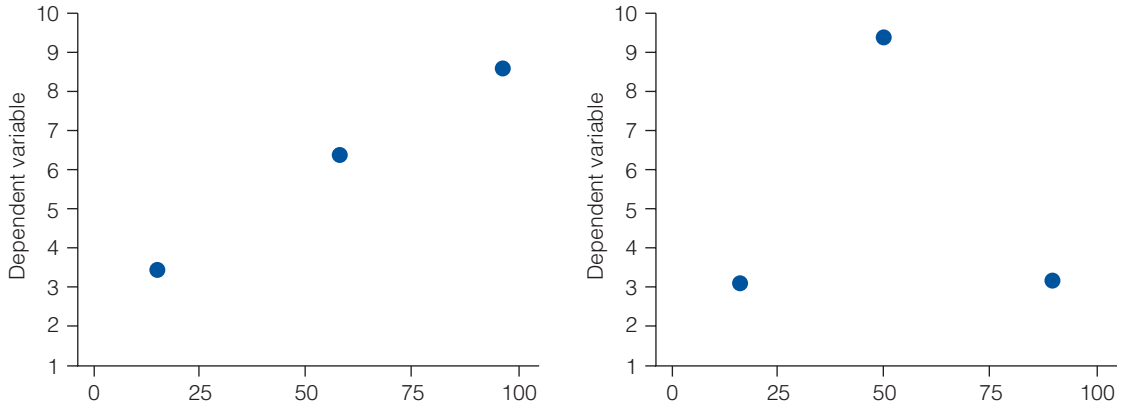

FIGURE 11.4 Having Three Levels of the Independent Variable (Three Data Points) Helps You Identify the Shape of the Functional Relationship

With these three points, we can be relatively confident that the relationship is linear (fits a straight line). Most nonlinear relationships (see Figure 11.3) would not 1produce data that would fit these three data points.

If we had these three data points, we could be relatively confident that the relationship is curvilinear (fits a curved line). Specifically, we would suspect that we had a quadratic relationship: a relationship shaped like a "U" or an upside-down "U."

In that case, you could confidently predict that 70 minutes of exercise would be less beneficial for increasing happiness than 100 minutes of exercise.

If, on the other hand, the relationship was S-shaped (as in Figure 11.3), you might get the following pattern of results:

Group 1	0 minutes of exercise per day	0.0 self-rating of happiness
Group 2	50 minutes of exercise per day	10.0 self-rating of happiness
Group 3	100 minutes of exercise per day	10.0 self-rating of happiness

In that case, you would predict that a person who exercised 70 minutes would do as well as someone exercising 100 minutes a day.

The more groups you use, the more accurately you can pin down the shape of the functional relationship. Yet, despite this fact, you do not need to use numerous levels of the independent variable. Why? Because nature prefers simple patterns. That is, most functional relationships are linear (straight lines), and few are more complex than U-shaped functions. Consequently, you will rarely need more than four levels of the independent variable to pin down a functional relationship. In fact, you will usually need no more than three carefully chosen levels.

Conclusions: Multilevel Experiments and External Validity

In summary, knowing the functional relationship between two variables is almost as important as knowing that a relationship exists. If you want to give practical advice, you should be able to say more than: "If you exercise 100 minutes a day, you will be happier than someone who exercises 0 minutes a day." Who exercises exactly 100 minutes a day? You want to be able to generalize your results so that you can tell people the effects of exercising 50 minutes, 56 minutes, 75 minutes, and so forth. Yet, you have no intention of testing the effects of every possible amount of exercise a person might do. Instead, you want to test only a handful of exercise levels. If you choose these levels carefully, you will be able to map the functional relationship between the variables. Mapping the functional relationship, in turn, will allow you to make educated predictions about the effects of treatment levels that you have not directly tested.

When applying psychology, you need to know the functional relationship so you can know how much of a therapy or other treatment to administer. How much is too little? At what point is additional treatment not worth it? How much is too much? If you know the answers to these questions, not only do you avoid wasting your time and your client's time on unnecessary treatments, but you also free up time and resources to help a client who needs it (Tashiro & Mortensen, 2006).

When mapping functional relationships, psychologists often manipulate independent variables that have names starting with "number of," such as number of others, number of milligrams of a drug, or number of seconds of exposure to a stimulus. You may be inspired by studies like the following classics.

- In Asch's (1955) line-judging experiments, he led participants to believe that they were part of a group that was participating in a visual perception study. The participant's job was to pick the line on the right that matched the line on the left. In reality, the experiment was a social influence experiment, and the other members of the group were really confederates (assistants) of the experimenter. Asch wanted to know whether the size of group would affect people's conformity to the group. He found that as group size went from two to five, participants were more likely to conform. However, he found that increasing the group size beyond seven actually decreased the chances that participants would go along with the group.
- In Latané, Williams, and Harkins's (1979) social loafing experiments, investigators wanted to know how loafing would change as a function of group size. They found that adding two members to a group increased loafing, but that adding two members increased loafing more in smaller groups than in larger groups.
- In Milgram, Bickman, and Berkowitz's (1969) conformity experiment, confederates looked up at the sixth floor window of an office building. Because the researchers were interested in the effects of group size on conformity, the researchers had 1, 2, 3, 5, 10, or 15 confederates look up at the office building. Then, they counted the number of people passing by who also looked up. They found that the bigger their initial group, the stronger the group's influence.

- In Darley and Latané's (1968) study of helping behavior, participants thought they were talking via intercom to either one, two, or five other participants (actually, the participant was alone—the voices came from a tape-recording) when one of them had a seizure. They found that the *more* people participants thought were in the group, the *less* likely participants were to help.
- In Middlemist, Knowles, and Matter's (1976) urinal study, researchers found that the closer a confederate was standing to a participant, the longer it took for the participant to begin urinating.
- In Ambady and Rosenthal's (1993) "thin slices" experiments, participants watched—with the sound off—three video clips of a professor. The clips varied in length: One group saw 2-second clips, a second group saw 5-second clips, and a third group saw 10-second clips. The researchers found that participants in all three groups gave the professor the same ratings as students who sat in the professor's class all term gave that professor.
- In Basson, McInnes, Smith, Hodgson, and Koppiker's (2002) study of the effect of Viagra on women, neither the 10-mg, 50-mg, or 100-mg doses of Viagra were more effective in increasing sexual response than a placebo.

Although it is easy to map functional relationships when the name of your independent variable starts with "number of," realize that you can map functional relationships between most variables because—with a little work—most variables can be quantified. If you can manipulate a variable between two extreme levels (e.g., low and high), you can probably also manipulate it in between those extremes (e.g., medium). To illustrate, consider a variable that is not obviously quantitative: similarity. Byrne (1961) manipulated similarity from 0% to 100% by (a) making participants believe they were seeing another student's responses to an attitude survey and then (b) varying the proportion of responses that matched the participant's attitudes from 0% to 100%.

If your independent variable involves exposing participants to a stimulus, you can usually quantify your manipulation by doing some work before you start your study. Specifically, you could (a) produce several variations of the stimulus, (b) have volunteers rate each of those variations, and then (c) use the variations that have the values you want in your experiment. For example, suppose you wanted to manipulate physical attractiveness by showing participants photos of people who varied in attractiveness. You could take a photo of an attractive person, then (a) get some less attractive photos of the person by messing with the person's makeup or by messing with their picture using a computerized photo editing program, (b) have some volunteers rate the attractiveness of each photo on a 0-to-10 scale, and (c) use, as your three stimuli, the photos that were consistently rated 4, 6, and 8, respectively. Realize that this scaling strategy is not just for pictures: If your manipulation was "severity of a crime" or "legitimacy of an excuse" or almost anything else, you could still use this scaling strategy.

Even without scaling your independent variable, you may still be able to order the levels of your manipulation from least to most and then see whether more of the manipulation creates more of an effect. For example, in one study

(Risen & Gilovich, 2007) all participants were to imagine the following scenario:

> You bought a lottery ticket from a student who was organizing a lottery. However, on the day of the drawing, you left your money at home and so had no money to buy lunch. To buy lunch, you sold your ticket back to the student who sold it to you.

Then, 1/3 of the participants were told to imagine that the lottery ticket was eventually purchased by their best friend, 1/3 were told to imagine that the ticket was purchased by a stranger, and 1/3 were told to imagine that the ticket was purchased by their "ex-friend and least favorite person at school" (p. 16). As Risen and Gilovich predicted, the *less* the participants liked the person who would eventually own their original ticket, the *more* likely participants thought that ticket would be the winning ticket. Conversely, Young, Nussbaum, and Monin (2007) showed that if a disease could be spread by sexual contact, people were reluctant to have themselves tested for it, and this reluctance was unaffected by whether sexual contagion was an unlikely (e.g., only 5% of the cases were due to sexual contact) or likely (e.g., 90% of the cases were due to sexual contact) cause of the disease.

Using Multiple Groups to Improve Construct Validity

You have seen that multilevel experiments—because their results can generalize to a wider range of treatment levels—can have more external validity than simple experiments. In this section, you will learn that multilevel experiments can also have more construct validity than simple experiments.

Confounding Variables in the Simple Experiment

In Chapter 10, you saw that thanks to random assignment, simple experiments are able to rule out the effects of variables unrelated to the treatment manipulation. For example, because of random assignment, the effects of participant variables such as gender, race, and personality usually will not be confused for a treatment effect. In other words, a statistically significant difference between the control group and the experimental group at the end of the experiment will probably not be due to the groups being different before the treatment was introduced.

So, simple experiments effectively control for the effects of variables that have nothing to do with the treatment manipulation. But what if the treatment is manipulating more than the one variable it's supposed to be manipulating? For instance, what if an "exercise" manipulation is also manipulating social support? Simple experiments are often unable to rule out the effects of variables that are manipulated along with the treatment.

In an ideal world, this limitation of the simple experiment would not be a problem. Your treatment would be a pure manipulation that creates one—and only one—systematic difference between the experimental group and the control group. Unfortunately, it is rare to have a perfect manipulation. Instead, the treatment manipulation usually produces several differences between how the experimental and control groups are treated.

For example, suppose that a simple experiment apparently found that the "attractive" defendant was more likely to get a light sentence than the "unattractive" defendant. We would know that the "attractiveness" manipulation

had an effect. However, it could be that in addition to manipulating attractiveness, the researchers also manipulated perceived wealth. Thus, wealth, rather than attractiveness, might account for the manipulation's effect. Specifically, people may be less likely to give wealthy defendants long sentences.

Because of impurities in manipulations, you often end up knowing that the treatment manipulation had an effect, but not knowing whether the treatment had an effect because it manipulated (a) the variable you wanted to manipulate or (b) some other variable that you did not want to manipulate (the impurity). In short, simple experiments may lack construct validity because the independent variable manipulation is contaminated by variables that are unintentionally manipulated along with the treatment. In technical terminology, the manipulation's construct validity is weakened by **confounding variables**: variables, other than the independent variable, that may be responsible for the differences between your conditions.

The following example[1] illustrates the general problem of confounded manipulations. Imagine being in a classroom that has five light switches, and you want to know what the middle light switch does. Assume that in the "control" condition, all the light switches are off. In the "experimental" condition, you want to flick the middle switch. However, because it is dark, you accidentally flick on the middle three switches. As the lights come on, the janitor bursts into the room, and your "experiment" is finished. What can you conclude?

You can conclude that your manipulation of the light switches had an effect. That is, your study has internal validity. But, because you manipulated more than just the middle light switch, you can't say that you know what the middle light switch did. Put another way, if you were to call your manipulation a "manipulation of the middle switch," your manipulation would lack construct validity.

Because of confounding variables, it is often hard to know what it is about the treatment that caused the effect. In real life, variables are often confounded. For example, your friend may know she got a hangover from drinking too much wine, but not know whether it was the alcohol in the wine, the preservatives in the wine, or something else about the wine that produced the awful sensations. A few years ago, a couple of our students joked that they could easily test the hypothesis that alcohol was responsible. All they needed us to do was donate enough money to buy mass quantities of a pure manipulation of alcohol—180 proof, totally devoid of impurities. These students understood how confounding variables can contaminate real-life manipulations—and how confounding variables can make it hard to know what it was about the manipulation that caused the effect.

Having a multiple-group experiment can allow you to know what it is about the source that causes a treatment's effect. For example, if you wanted to look at the effects of cell phones on driving behavior, you could have a no cell phone group, a cell phone group, and a cell phone with headset group. By comparing the regular cell phone group to the headset group, you might be able to see whether reaching for the phone was a source of the cell phone users' driving problems (Strayer & Drews, 2008). To see how having more

[1] We are indebted to an anonymous reviewer for this example and other advice about confounding variables.

than two groups has helped researchers track down the source of a treatment's effect, consider the following examples.

- Gesn and Ickes (1999) found that participants who saw a video of another person did a passable job at knowing what that person was thinking. But why? Was it the person's words—or was it their nonverbal signals? To find out, Gesn and Ickes compared one group that heard only the words (audio only) to another group that got only the nonverbal signals. (The nonverbal group saw video of the person accompanied by a filtered sound track that allowed participants to hear the pitch, loudness, and rhythm of the person's speech, but not the actual words.) Gesn and Ickes found that the words, rather than nonverbal signals, were what helped participants figure out what the person was thinking. Specifically, whereas the audio-only group did nearly as well as the normal video group, the video with filtered audio group did very poorly.
- Langer, Blank, and Chanowitz (1978) had their assistant get into lines to use the copier and then ask one of three questions:

  1. Can I cut in front of you?
  2. Can I cut in front of you because I'm in a rush?
  3. Can I cut in front of you because I want to make a copy?

  The researchers found that 60% of participants in the no-excuse condition let the assistants cut in front, 94% of the participants in the good-excuse condition let the assistants cut in, and 93% of the participants in the poor-excuse condition let the assistants cut in front of them. By having both a no-excuse control group and a good-excuse control group, the researchers were able to establish that it was (a) important to have an excuse but (b) the quality of the excuse was unimportant.

- In the false memory study we discussed earlier, Loftus (1975) included a control group who, like the experimental group, was asked questions about objects that weren't in the film, but who, unlike the experimental group, were not asked questions that implied that those objects were in the film (e.g., the control group might be asked "Did you see a red stop sign?" whereas the experimental group would be asked, "Did you see the red stop sign?"). The fact that this control group did not have false memories allowed Loftus to discover that the false memories in the leading question condition were caused by suggesting that the object was present—and not by the mere mention of the false object.
- Lee, Frederick, and Ariely (2006) found that people told that they were about to drink some beer that had vinegar added to it rated the beer more negatively than participants not told about the vinegar. One possibility for this finding is that participants merely obeyed demand characteristics: Participants might expect that the experimenter wanted them to give low ratings to vinegar-tainted beer. Fortunately, Lee et al. were able to rule out this possibility because they had a control group that was told about the vinegar *after* tasting the beer—and that "after" group rated the beer as positively as the group that didn't know about the vinegar. Consequently, the researchers were able to conclude that knowing about the vinegar *beforehand* changed how the beer tasted to participants.

- Baumeister, DeWall, Ciarocco, and Twenge (2005) found that participants believing they would spend the future alone exhibited less self-control than participants believing they would spend the future with friends. However, this finding could mean either that social rejection leads to less self-control or that expecting unpleasant outcomes leads to less self-control. Therefore, Baumeister et al. added a control group of participants who were led to expect an unpleasant, injury-riddled future. That "misfortune" group did not experience a loss of self-control, suggesting that it was rejection, not negative events, that caused the lowered self-control.

To understand how confounding variables can contaminate a simple experiment, let's go back to the simple experiment on the effects of exercise that we proposed earlier in this chapter. You will recall that the experimental group got 100 minutes of exercise class per day, whereas the control group got nothing. Clearly, the experimental group participants were treated differently from the control group participants. The groups didn't differ merely in terms of the independent variable (exercise). They also differed in terms of several other (confounding) variables: The exercise group received more attention and had more structured social activities than the control group.

Hypothesis-Guessing in Simple Experiments. Furthermore, participants in the experimental group knew they were getting a treatment, whereas participants in the control group knew they were not receiving any special treatment. If experimental group participants suspected that the exercise program should have an effect, the exercise program may appear to have an effect—even if exercise does not really improve mood. In other words, the construct validity of the study might be ruined because the experimental group participants guessed the hypothesis (**hypothesis-guessing**).

Because of the impurities (confounding variables) of this exercise manipulation, you cannot say that the difference between groups is due to exercise by itself. Although all manipulations have impurities, this study's most obvious—and avoidable—impurities stem from having an **empty control group**: a group that gets no treatment, not even a placebo (a placebo is a treatment that doesn't have an effect, other than possibly by changing a participants' expectations). Thus, if you chose to use a placebo control group instead of the empty control group, you could reduce the impact of confounding variables.

Increasing Validity Through Multiple Control Groups

Choosing the placebo control group over the empty control group does, however, often come at a cost. Often, it would be better to have both control groups.

To see how hard it can be to choose between an empty control group and a placebo group, consider the studies comparing the effect of antidepressant drugs to the effect of a placebo. If those simple experiments had compared groups getting antidepressants to empty control groups, those studies would have grossly overestimated the effectiveness of antidepressant drugs (Kirsch, Moore, Scoboria, & Nicholls, 2002). However, because those studies did not use empty control groups, they don't tell us the difference between getting the

drug and receiving no treatment. Given that patients will be choosing between drug treatment and no drugs (Moerman, 2002), the lack of an empty control group is a problem. It would have been nice to have compared the antidepressant group to both an empty control group as well as to a placebo group.

The Value of a Placebo Group. To take another example of the difficulty of choosing between a placebo group and an empty control group, let's go back to the problem of examining the effects of exercise on mood. If you use an empty control group that has nothing done to its participants, interpreting your results may be difficult. More specifically, if the exercise group does better than this "left alone" group, the results could be due to hypothesis-guessing (e.g., participants in the exercise condition figuring out that exercise should boost their mood) or to any number of confounding variables (such as socializing with other students in the class, being put into a structured routine, etc.).

If, on the other hand, you use a placebo-treatment group (for example, meditation classes), you would control for some confounding variables. For example, both your treatment and placebo groups would be assigned to a structured routine. Now, however, your problem is that you only know how the treatment compares to the placebo: You do not know how it compares to no treatment. Consequently, you won't know what the treatment's effect is.

The Value of an Empty Control Group: "Placebos" May Not Be Placebos. You won't know what the effect of your treatment is because you do not know what the effect of your placebo treatment is. Ideally, you would like to believe that your placebo treatment has no effect. In that case, if the treatment group does worse than the placebo group, the treatment is harmful; if the treatment group does better, the treatment is helpful.

If, however, what you hope is a purely placebo treatment turns out to be a treatment that really does have an effect, you are going to have trouble evaluating the effect of your treatment. For example, suppose you find that the exercise group is more depressed than the meditation group. Could you conclude that exercise increases depression? No, because it might be that although exercise reduces depression, meditation reduces it more. Conversely, if you found that the exercise group is less depressed than the meditation group, you could not automatically conclude that exercise decreases depression. It may be that meditation increases depression greatly, and exercise increases depression only moderately: Exercise may merely be the lesser of two evils.

To find out whether exercise increases or decreases depression, you need to compare the exercise group to a no-treatment group. Thus, if you were interested in the effects of exercise on depression, you have two options: (1) Use a simple experiment and make the hard choice between an empty control group and a placebo group, or (2) use a multiple-group experiment so that you can include both an empty and a placebo control group.

Using Multiple Imperfect Control Groups to Compensate for Not Having the Perfect Control Group. Even if you are sure you do not want to use an empty control group, you may still need more than one control group because you

will probably not have the perfect control group. Instead, you may have several groups, each of which controls for some confounding variables but not for others. If you were to do a simple experiment, you may have to decide which of several control groups to use. Choosing one control group—when you realize you need more than one—is frustrating. It would be better to be able to use as many as you need.

But how often do you need more than one control group? More often than you might think. In fact, even professional psychologists sometimes underestimate the need for control groups. Indeed, many professional researchers get their research articles rejected because a reviewer concluded that they failed to include enough good control groups (Fiske & Fogg, 1990).

You often need more than one control group so that your study will have adequate construct validity. Even with a poor control group, your study has internal validity: You know that the treatment group scored differently than the control group. But what is it about the treatment that is causing the effect? Without good control group(s), you may think that one aspect of your treatment (the exercise) is causing the effect, when the difference is really due to some other aspect of your treatment (the socializing that occurs during exercise).

To illustrate how even a good control group may still differ from the experimental group in several ways having nothing to do with the independent variable, consider the meditation control group. The meditation control group has several advantages over the empty control group. For example, if the exercise group was less depressed than a meditation control group, we could be confident that this difference was not due to hypothesis-guessing, engaging in structured activities, or being distracted from worrisome thoughts for awhile. Both groups received a "treatment," both engaged in structured activities, and both were distracted for the same length of time.

The groups, however, may differ in that the exercise group did a more social type of activity, listened to louder and more upbeat music, and interacted with a more energetic and enthusiastic instructor. Therefore, the exercise group may be in a better mood for at least three reasons having nothing to do with exercise: (1) the social interaction with their exercise partners, (2) the upbeat music, and (3) the upbeat instructor.

To rule out all these possibilities, you might use several control groups. For instance, to control for the "social activity" and the "energetic model" explanations, you might add a group that went to a no-credit acting class taught by an enthusiastic professor. To control for the music explanation, you might add a control group that listened to music or perhaps even watched aerobic dance videos. By using all of these control groups, you may be able to rule out the effects of confounding variables.

ANALYZING DATA FROM MULTIPLE-GROUP EXPERIMENTS

You have just learned that multiple control groups may give you more construct validity than one control group. Earlier, you learned that multiple treatment groups allow you to more accurately map the functional relationship between the independent variable and the dependent variable than a two-group experiment. Before that, you learned that the multiple-group

experiment allows you to compare more treatments at one time than a two-group experiment. In short, you have learned that there are at least three good reasons to conduct a multiple-group experiment instead of a simple experiment:

1. to improve construct validity
2. to map functional relationships
3. to compare several treatments at once

However, before you conduct a multiple-group experiment, you should understand how it will be analyzed because the way that it will be analyzed has implications for (a) what treatment groups you should use, (b) how many participants you should have, and even (c) what your hypothesis should be.

Even if you never conduct a multiple-group experiment, you will read articles that report results of such experiments. To understand those articles, you must understand the logic and vocabulary used in analyzing them.

Analyzing Results From the Multiple-Group Experiment: An Intuitive Overview

As a first step to understanding how to analyze the results of multiple-group experiments, let's look at data from three experiments that compared the effects of no-treatment, meditation, and aerobic exercise on happiness. All of these experiments had 12 participants rate their feelings of happiness on a 0-to-100 (not at all happy to very happy) scale. Here are the results of Experiment A:

	NO-TREATMENT	MEDITATION	EXERCISE
	50	51	53
	51	53	53
	52	52	54
	51	52	52
Group Means	51	52	53

Compare these results to the results of Experiment B:

	NO-TREATMENT	MEDITATION	EXERCISE
	40	60	78
	42	60	82
	38	58	80
	40	62	80
Group Means	40	60	80

Are you more confident that Experiment A or Experiment B found a significant effect for the treatment variable? If you say B, why do you give B as your answer? You answer B because there is a *bigger difference between the groups* in Experiment B than in Experiment A. That is, the group means for Experiment B are further apart than the group means for Experiment A. Group B's means being further apart—what statisticians call greater **variability between group means**—lead you to think that Experiment B is more likely to be the study that obtained significant results for two reasons.

First, you intuitively realize that to find a treatment effect, you need between-group variability. After all, if the between-group variability was zero (indicating that the means of the exercise group, the no-treatment group, and the meditation group were all the same), you couldn't argue that the treatment had an effect.

Second, you intuitively realize a small difference between group means might easily be due to chance (rather than to the treatment), but a larger difference is less likely to be due to chance.[2] Thus, you realize that the more variability there is between group means, the more likely it is that at least some of that variability is due to treatment.

Now, compare Experiment B with Experiment C. Here are the results of Experiment C:

	EXERCISE	NO-TREATMENT	MEDITATION
	10	10	100
	80	90	80
	60	60	60
	10	80	80
Group Means	40	60	80

Which experiment do you think provides stronger evidence of a treatment effect—Experiment B or Experiment C? Both experiments have the same amount of variability between group means. Therefore, unlike in our first example, you cannot use the rule of choosing the experiment with the means that differ the most to choose Experiment B. Yet, once again, you will pick Experiment B. Why?

You will pick Experiment B because you are concerned about one aspect of Experiment C: the extreme amount of variability within each group. You realize the only reason scores within a group vary is random error. (If participants in the same treatment group get different scores, those different scores can't be due to the treatment. Instead, the differences in scores must be due to nontreatment variables, such as individual differences. In a randomized experiment, such nontreatment variables become random error.) Thus, you see that Experiment C is more affected by random error than Experiment B.

[2] Similarly, if your favorite team lost by one point, you might blame luck. However, if your team lost by 30 points, you would be less likely to say that bad luck alone was responsible for the defeat.

The large amount of random error in Experiment C (as revealed by the *within-groups* variability) bothers you because you realize that this random error—rather than the treatment—might be the reason the groups differ from one another. That is, the same random variability that makes individual scores within a group differ from each other might also make the group means differ from each other.[3] In Experiment B, on the other hand, the small amount of within-group variability indicates that there is virtually no random variability in the data. Therefore, in Experiment B, you feel fairly confident that random error is *not* causing the group means to differ from one another. Instead, you believe that the means differ from one another because of the treatment.

Intuitively then, you understand the three most important principles behind analyzing the results of a multiple-group experiment. Specifically, you realize the following:

1. Within-groups variability is not due to the treatment, but instead is due to random error. That is, differences within a treatment group can't be due to the treatment because everyone in the group is getting the same treatment. Instead, differences among group members must be due to random factors such as individual differences and random measurement error.
2. Between-groups variability is not a pure measure of treatment effects. Admittedly, if the treatment has an effect, the means of groups getting different levels of treatment should differ from one another. However, even if the treatment has no effect, the group means will probably still differ from one another because of random error. Thus, between-group variability is affected by **both** random error and treatment effects.
3. If you compare between-group variability (the effects of random error plus any treatment effects) to within-group variability (the effects of random error alone), you may be able to determine whether the treatment had an effect.

Analyzing Results From the Multiple-Group Experiment: A Closer Look

You now have a general idea of how to analyze data from a multiple-group study. To better understand the logic and vocabulary used in these analyses—a must if you are to understand an author's or a computer's report of such an analysis—read the next few sections.

Within-Groups Variability: A Pure Measure of Error

As you already know, within-groups variability does not reflect the effects of treatment. Instead, it reflects the effects of random error. For example, because all the participants in the meditation group are getting the same

[3] To get a sense of how random sampling error might cause the group means to differ, randomly sample two scores from the no-treatment group (scores are in the table on page 400). Compute the mean of this group. If you do this several times, you will get different means. These different means can't be due to a treatment effect because none of the participants in any of your samples are receiving the treatment. The reason you are getting different means even though you are sampling the same group is random sampling error. Fortunately, statistics can help us determine how likely it is that the differences among group means are entirely due to random error.

treatment (meditation), any differences among those participants' scores can't be due to the treatment. Instead, the differences among scores of meditation group participants are due to such random factors as individual differences, unreliability of the measure, and lack of standardization. Similarly, differences among the scores of participants in the no-treatment group are due not to treatment, but to irrelevant random factors. The same is true for differences within the exercise group. Thus, calculating within-groups variability will tell us the extent to which chance causes individual scores to differ from each other.

To measure this within-groups variability, we first look at the variability of the scores within each group. To be more specific, we calculate an index of variability called the variance. If we have three groups, we could calculate the variance within each group. Each of these three within-group variances would be an estimate of the extent to which the groups could differ due to random error alone. However, we do not need three different estimates of random error—we just need one good one. To end up with one estimate of variability due to random error, we average all three within-group variances to come up with the best estimate of random variability—the within-groups variance.

Fortunately, we can use this estimate of how much random error causes individual scores to differ from each other to estimate the extent to which random error is likely to cause group means to differ from each other. Partly because this **within-groups variance** gives us an index of the degree to which *random error* alone may cause your group means to differ, within-groups variance is often referred to as **error variance**.

Between-Groups Variability: Error Plus (Possibly) Treatment

Once you have an index of the degree to which your groups could vary from each other due to chance alone (the within-groups variance), the next step is to get an index of the degree to which your groups actually vary from one another. It is at this step where it becomes clear that you cannot use a *t* test to analyze data from a multiple-group experiment. When using a *t* test, you determine the degree to which the groups differ from one another in a straightforward manner: You subtract the average score of Group 1 from the average score of Group 2. Subtraction works well when you want to compare two groups, but it does not work well when you have more than two groups because you can subtract only two scores at a time. So, if you have three groups, which two groups do you compare? Group 1 with Group 2? Or, Group 2 with Group 3? Or, Group 1 with Group 3?

You might answer this question by saying "all of the above." You are saying that, with three groups, you would do three *t* tests: one comparing Group 1 against Group 2, a second comparing Group 1 against Group 3, and a third comparing Group 2 against Group 3. However, that's not allowed!

An analogy will help you understand why you cannot use multiple *t* tests. Suppose a stranger comes up to you with a proposition: "Let's bet on coin flips. If I get a 'head,' you give me a dollar. If I don't, I give you a dollar." You agree. He then proceeds to flip three coins at once and then makes you pay up if even one of the coins comes up heads. Why is this unfair? This is unfair because he misled you: You thought he was going to flip only one

coin at a time, so you thought he had only a 50% chance of winning. But because he's flipping three coins at a time, his chances of getting at least one head are much better than 50%.[4]

When you do multiple *t* tests, you are doing basically the same thing as the coin hustler. You start by telling people the odds that a single *t* test will be significant due to chance alone. For example, if you use conventional significance levels, you would tell people that if the treatment has no effect, the odds of getting a statistically significant result for a particular *t* test are less than 5 in 100. In other words, you are claiming that your chance of making a Type 1 error is no more than 5%.

Then, just as the hustler gave himself more than a 50% chance of winning by flipping more than one coin, you give yourself a more than 5% chance of getting a statistically significant result by doing more than one *t* test. The 5% odds you quoted would hold only if you had done a single *t* test. If you are using *t* tests to compare three groups, you will do three *t* tests, which means the odds of at least one turning out significant by chance alone are much more than 5%.[5]

So far, we've talked about the problems of using a *t* test when you have a three-group experiment. What happens if your experiment has more than three groups? Then, the *t* test becomes even more deceptive (just as the coin hustler would be cheating even more if he flipped more than three coins at a time). The more groups you use in your experiment, the greater the difference between the significance level you report and the actual odds of at least one *t* test being significant by chance (Hays, 1981).

To give you an idea of how great the difference between your stated significance level and the actual odds can be, suppose you had six levels of the independent variable. To compare all six groups with one another, you would need to do 15 *t* tests. If you did that and used a .05 significance level, the probability of getting at least one significant effect by chance alone would be more than 50%: Your risk of making a Type 1 error would be 10 times greater than you were claiming it was!

As you have seen, the *t* test is not useful for analyzing data from the multiple-group experiment because it measures the degree to which groups differ (vary) by using subtraction—and you can only subtract two group averages at a time. To calculate the degree to which more than two group means vary, you need to calculate a variance between those means.

The between-groups variance indicates the extent to which the group means vary (differ). Thus, if all your groups have the same mean, between-groups variance would be zero because there would be no (zero) differences between your group means. If, on the other hand, there are large differences between the group means, between-group variance will be large.

So, the size of the between-groups variance depends on the extent to which the group means differ. But what affects the extent to which the group means differ? As you saw earlier, there are two factors.

One factor is random error. Even if the treatment has no effect, random error alone will almost always cause differences between the group means.

---

[4] To be more precise, his chances of getting at least one head are 87.5%.

[5] To be more precise, your chances are 14.26%.

If the experiment uses an unreliable measure, few participants, and poorly standardized procedures, random error alone may cause large differences between the group means. If the experiment uses a reliable measure, many participants, and highly standardized procedures, random error alone would tend to cause smaller differences between the group means. In short, when there is no treatment effect, the groups will still differ from each other due to random error. To be more specific, when there is no treatment effect, between-groups variance should be roughly equivalent to a more direct measure of random error: within-groups variance.

The other factor that *may* affect the extent to which the groups differ from each other is the treatment effect. If the treatment has an effect, the differences between the group means should be greater than when the treatment doesn't have an effect. Because of the treatment effect's influence on the size of the between-groups variance, the between-groups variance is often called **treatment variance**.

To recap, when there is a treatment effect, the between-group variance is the sum of two quantities: an estimate of random error plus an estimate of treatment effects. Therefore, if the treatment has an effect, between-groups variance (which is affected by the treatment plus random error) will be larger than the within-groups variance (which is affected only by random error).

Comparing Between-Groups Variance to Within-Groups Variance: Are the Differences Between Groups Due to More Than Random Error?

Once you have the between-groups variance (an estimate of random error plus any treatment effects) and the within-groups variance (an estimate of random error), the next step is to compare the two variances. If the between-groups variance is larger than the within-groups variance, some of the between-groups variance may be due to a treatment effect. The statistical *analysis* that allows you to compare the between-groups *variance* to the within-groups *variance* and thereby determine whether the treatment had an effect is called **analysis of variance (ANOVA)**.

When doing an ANOVA, you compare two variances by dividing the between-groups variance by the within-groups variance. That is, you set up the following ratio:

$$\frac{\text{Between-Groups Variance}}{\text{Within-Groups Variance}}$$

Instead of using the term *variance*, you are more likely to see the term *mean square*. Thus, you are more likely to read about authors setting up the following ratio:

$$\frac{\text{Mean Square Between Groups}}{\text{Mean Square Within Groups}}$$

Note that authors tend to leave off the word *groups*. As a result, you are likely to see the ratio described as

$$\frac{\text{Mean Square Between}}{\text{Mean Square Within}}$$

To shorten the expression even further, authors tend to abbreviate *M*ean *S*quare as *MS*, *M*ean *S*quare *B*etween as *MSB,* and *M*ean *S*quare *W*ithin as *MSW*. Therefore, you are likely to see the ratio of the variances described as

$$\frac{MSB}{MSW}$$

To complicate things further, authors may not use the terms *between* or *within.* Rather than use a name that refers to how these variances were calculated (looking at differences *between* group means for MS *between* and looking at differences *within* groups for *MS within)*, authors may instead use a name that refers to what these variances estimate. Thus, because between-groups variance is, in part, an estimate of treatment effects, authors may refer to mean square between as mean square *treatment* (abbreviated *MST*). Similarly, because within-groups variance is an estimate of the degree to which random error is affecting estimates of the treatment group means, authors may refer to mean square *within* as mean square *error* (abbreviated *MSE*).

Regardless of what names or abbreviations authors give the two variances, the ratio of the between-groups variance to the within-groups variance is called the **F ratio**. Consequently, the following three ratios are all *F* ratios:

$$\frac{MSB}{MSW} = \frac{MS\,\text{Treatment}}{MS\,\text{Error}} = \frac{MST}{MSE}$$

In ANOVA summary tables, terms are shortened even more. Thus, when scanning computer printouts or when reading articles, you may see tables resembling the one below:

SOURCE	MEAN SQUARE	*F*
Treatment	10	2
Error	5	

Why an F of 1 Does Not Show That the Treatment Had an Effect. Conceptually, the *F* ratio can be portrayed as follows:

$$F = \frac{\text{Random Error} + \text{Possible Treatment Effect}}{\text{Random Error}}$$

By examining this conceptual formula, you can see that the *F* ratio will rarely be much less than 1. To illustrate, imagine that the null hypothesis is true: There is no (zero) treatment effect. In that case, the formula is (random error + 0)/random error, which reduces to random error/random error. As you know, if you divide a number by itself (e.g., 5/5, 8/8), you get 1.[6]

[6] The only exception is that 0/0 = 0.

You now know that if the null hypothesis were true, the *F* ratio would be approximately 1.00.[7] That is,

$$F = \frac{\text{Random Error} + 0}{\text{Random Error}} = \frac{\text{Random Error}}{\text{Random Error}} = 1.00$$

But what would happen to the *F* ratio if the treatment had an effect? To answer this question, let's look at what a treatment effect would do to the top and the bottom half of the *F* ratio.

If the treatment has an effect, the top of the *F* ratio—the between-groups variance—should get bigger. Not only is the between-groups variance affected by random error (as it was when the treatment did not have an effect), but now that the treatment is also making the group means differ, between-groups variance is also influenced by the treatment.

We just explained that a treatment effect increases the *top* of the *F* ratio, but what does a treatment effect do to the *bottom* of the *F* ratio? Nothing. Regardless of whether there is a treatment effect, the bottom of the *F* ratio, the within-groups variance, always represents only random error: With or without a treatment effect, a group's scores differ from one another solely because of random error.

Let's now use our knowledge of how treatment effects influence the top and bottom parts of the *F* ratio to understand how treatment effects influence the entire *F* ratio. When there is a treatment effect, the differences between group means are due not only to random error (the only thing that affects within-groups variance) but also to the treatment's effect. Consequently, when there is a treatment effect, the between-groups variance (an index of random error plus treatment effect) should be larger than the within-groups variance (an index of random error alone). Put more mathematically, when there is a treatment effect, you would expect the ratio of between-groups variance to within-groups variance to be greater than 1. Specifically,

$$F = \frac{\text{Between-Groups Variance (Treatment + Random Error)}}{\text{Within-Groups Variance (Random Error)}} > 1,$$

when the treatment has an effect.

Using an F Table. Not all *F*s above 1.00 are statistically significant, however. To determine whether an F ratio is enough above 1.00 to indicate that there is a significant difference between your groups, you need to consult an *F* table, like the one in Appendix F.

Calculating Degrees of Freedom. To use the *F* table, you need to know two degrees of freedom: one for the top of the *F* ratio (between-groups variance,

[7] If you get an *F* below 1.00, it indicates that you have found no evidence of a treatment effect. Indeed, in the literature, you will often find statements such as, "There were no other significant results, all *F*s < 1." If you get an *F* substantially below 1.00, you may want to check to be sure you did not make a computational error. If your *F* is negative, you have made a computational error: *F* can't be less than 0.

TABLE **11.1**
Calculating Degrees Of Freedom

SOURCE OF VARIANCE (*SV*)	CALCULATION OF *DF*
Treatment (between groups)	Number of Groups–1 (*G*–1)
Within subjects (error variance)	Number of participants minus number of groups (*N*–*G*)
Total	*N*–1

MS treatment) and one for the bottom of the *F* ratio (within-groups variance, *MS* error).

Calculating the degrees of freedom for the top of the *F* ratio (between-groups variance) is simple. It's just one less than the number of values of the independent variable. So, if you have three values of the independent variable (no-treatment, meditation, and exercise), you have 2 (3–1) degrees of freedom. If you had four values of the independent variable (e.g., no-treatment, meditation, archery, aerobic exercise), you would have 3 (4–1) degrees of freedom. Thus, for the experiments we have discussed in this chapter, *the degrees of freedom for the between-groups variance equals the number of groups–1.*

Computing the degrees of freedom for the bottom of the *F* ratio (within-groups variance) is also easy. The formula is *N* (number of participants)–*G* (groups). Thus, if there are 20 participants and 2 groups, the degrees of freedom = 18 (20–2 = 18).[8]

Let's now apply this formula to some multiple-group experiments. If we have 33 participants and 3 groups, the *df* for the error term = 30 (because 33–3 = 30). If we had 30 participants and 5 groups, the *df* error would = 25 (because 30–5 = 25). To repeat, the simplest way of computing the error *df* for the experiments we discussed in this chapter is to use the formula *N*–G, where *N* = total number of participants and *G* = total number of groups (see Table 11.1).

Once you know the degrees of freedom, find the column in the $p < .05$ *F* table (Table 3 of Appendix F) that corresponds to those degrees of freedom. If your *F* ratio is larger than the value listed, the results are statistically significant at the $p < .05$ level.

Making Sense of an ANOVA Summary Table or Computer Printout. Usually, you will not have to look up *F* values in an *F* table. Instead, you will have a computer calculate *F* and look it up in a table for you. However, if you had a computer calculate *F* for you, you should make sure that the degrees of freedom on the printout are correct. If not, the computer has misunderstood your design or you have miscoded some data. If you had a computer calculate *F*

[8] As you may recall, you could have used this *N*–*G* formula to get the degrees of freedom for the *t* test described in Chapter 10. However, because the *t* test always compares 2 groups, people often memorize the formula *N*–2 for the *t* test instead of the more general formula *N*–*G*.

for you, the computer might provide you with an analysis of variance (ANOVA) summary table like this one:

SOURCE OF VARIANCE	SUM OF SQUARES (*SS*)	*df*	*MS*	*F*	*p*
Treatment (between)	88	2	44	44	<.05
Error (within)	12	12	1		
Total	100	14			

The first column, the source of variance column, may sometimes have only the heading "Source." The two main sources of variance will be your treatment (which may be labeled as "Treatment," "Between groups," "BG," "Groups," "Between," "Model," or the actual name of your independent variable) and random error (which may be labeled as "Error," "Within groups," "WG," or "Within").

The second column, the sum of squares column, may be labeled "Sum of Squares," "SS," or "Type III Sum of Squares." Note that if you add the sum of squares treatment to the sum of squares error, you will get the sum of squares total.

The third column, the degrees of freedom column, is often abbreviated *df*. As we mentioned earlier, you should check the *df* column to make sure that the analysis is based on the right number of treatment groups and the right number of participants. From the *df* column in our ANOVA table, we know two things. First, because the formula for the *df* treatment is G−1 and because the treatment *df* is 2, we know that a three-group ANOVA has been calculated (because 3 [groups]−1 = 2 [*df*]). Second, because the formula for total *df* is N−1 (number of participants−1) and because the total *df* is 14, we know that the ANOVA is based on data from 15 participants (because 15 [participants]−1 = 14 [*df*]).

The fourth column, the Mean Square column, is often abbreviated *MS*. The *MS* Treatment will be the *SS* Treatment divided by the *df* Treatment. Note that if the *MS* Treatment is *not* bigger than *MS* Error, the results will *not* be statistically significant.

The fifth column contains the *F* ratio. The *F* ratio is the *MS* Treatment divided by *MS* Error. In the table above *F* is 44 because 44 (*MST*) divided by 1 (*MSE*) = 44.

The sixth column, the *p* value column, tells you how likely it would be to get differences between the groups this large or larger if the null hypothesis (the null hypothesis is that the treatment has no effect) were true. In this case, *p* is less than .05, suggesting that it is unlikely that you would obtain these results if the null hypothesis were true. Traditionally, such results would be called "statistically significant." An author might start to summarize the results of such an ANOVA by writing, "Consistent with the hypothesis, the treatment had an effect, $F(2, 12) = 44$, $p < .05$."

The Meaning of Statistical Significance in ANOVA

If your results are statistically significant, what does that mean? *Statistical significance means that you can reject the null hypothesis*. In the multiple-group

experiment, the null hypothesis is that the differences among all your group means are due to chance. That is, all your groups are essentially the same. Rejecting this hypothesis means that, because of treatment effects, all your groups are *not* the same. In other words, you can conclude that at least two of your groups differ. However, such significant results raise two questions.

The first question is, "How large is the effect?" One way to get an estimate of the effect size is simply to look at the differences between the means. For example, looking at such differences suggests that the effect of antidepressants on relieving depression is only to increase scores by 2 points on a 50-point scale (Kirsch, Moore, Scoboria, & Nicholls, 2002). Another strategy is to compute **eta squared** (η^2): an estimate of effect size that ranges from 0 to 1 and is comparable to *r* squared.[9]

Computing eta squared from an ANOVA summary table is simple: Just divide the Sum of Squares Treatment by the Sum of Squares total. For example, in our ANOVA table, *SS* treatment was 88 and *SS* total was 100; therefore, eta squared was .88 (because 88/100 = .88)—indicating an extremely large effect. Thus, an author might start to describe such results by writing, "Consistent with the hypothesis, the treatment had an effect, $F(2, 12) = 44$, $p < .05$, $\eta^2 = 0.88$." Note that you would normally not get such a large eta squared. Indeed, social scientists tend to view any eta squared (or *r* squared) of .25 or above to be large (.09 to .25 is considered moderate; less than .09 is considered small).

The second question is, "Which groups differ from each other?" Even in a three-group experiment, there are several possibilities: Group 1 might differ from Group 2, and/or Group 2 might differ from Group 3, and/or Group 1 might differ from Group 3. As we just said, a significant *F* does not tell you which groups differ. Therefore, once you have performed an *F* test to determine that at least some of your groups differ, you need to do additional tests to determine which of your groups differ from one another.

Beyond ANOVA: Pinpointing a Significant Effect

You might think that all you would have to do to determine which groups differ is compare group means. Some group means, however, may differ from others solely due to chance. To determine which group differences are due to treatment effects, you need to do additional tests. These additional, more specific tests are called post hoc *t* tests.

Post Hoc *t* Tests Among Group Means: Which Groups Differ? At this point, you may be saying that you wanted to do *t* tests all along. Before you complain to us, please hear our two-pronged defense.

First, you can go in and do **post hoc tests** only *after* you get a significant *F*. That is, you can't legitimately use follow-up tests to ask "which of the groups differ" until you first establish that at least some of the groups do indeed differ. To do post hoc tests without finding a significant *F* is considered statistical malpractice. Such behavior would be like a physician doing a

[9] To learn about *r* squared, review our section titled "Coefficient of Determination" in Chapter 7, Box 10.2 in Chapter 10, or look at Appendix E.

specific test to find out which strain of hepatitis you had after doing a general test that was negative for hepatitis. At best, the test will not turn up anything, and your only problem will be the expense and pain of an unnecessary test. At worst, the test results will be misleading because the test is being used under the wrong circumstances. Consequently, you may end up being treated for a hepatitis you do not have. Analogously, a good researcher does not ask which groups differ from one another unless the more general, overall analysis of variance test has first established that at least some of the groups do indeed differ.[10]

Second, post hoc tests are not the same as conventional *t* tests. Unlike conventional *t* tests, post hoc *t* tests are designed to correct for the fact that you are doing more than two comparisons. As we mentioned earlier, doing more than one *t* test at the $p = .05$ level and claiming that you have only a 5% risk of making a *Type 1* error is like flipping more than one coin at a time and claiming that the odds of getting a "heads" are only 50%. In both cases, the odds of getting the result you hope for are much greater than the odds you are stating. Thus, we cannot simply do an ordinary *t* test. Instead, we must correct for the number of comparisons we are making. Post hoc *t* tests take into consideration how many tests are being done and make the necessary corrections.

At this point, we will not require you to know how to do post hoc tests. (If you want to know how to conduct a post hoc test, see Appendix F.) You should, however, be aware that if you choose to do a multiple-group experiment, you should be prepared to do post hoc analyses.

You should also be aware that if you read a journal article describing the results of a multiple-group experiment, you may read the results of post hoc tests. For example, you may read about a Bonferroni *t* test, Tukey test, Scheffe test, Dunnett test, Newman-Keuls test, Duncan, or an LSD test. When reading about the results of such tests, do not panic: The author is merely reporting the results of a post hoc test to determine which means differ from one another.

Post Hoc Trend Analysis: What Is the Shape of the Relationship? Rather than wanting to know which particular groups differ from one another, you may want to know the shape of the functional relationship between the independent and dependent variables so that you could either (a) better generalize to levels of the treatment that were not tested or (b) test a theory that predicts a certain functional relationship. If you want to know the shape of the functional relationship, instead of following up a significant main effect with post

---

[10] Although everyone agrees that you need to do an ANOVA before doing a post hoc *t* test, not everyone agrees that you need to do an ANOVA before doing other tests. Indeed, Robert Rosenthal (1992) argued that researchers should almost never do the general, overall *F* test. Instead, he argued that if you have specific predictions about which groups differ, you should do normal *t* tests to compare those group means. Those *t* tests are often called "planned comparisons" because the researcher planned to make those comparisons before collecting data. Planned comparisons involving t tests are sometimes also called "*a priori t* tests" ("a priori" means in advance) to emphasize that the *t* tests were done before peeking at the data. Sometimes, planned comparisons will be called "planned contrasts." One planned contrast that you will see when the researcher is trying to determine whether the two experimental groups differ from the control group or whether the two control groups differ from the experimental group is the "two vs. one" contrast.

hoc tests between group means, follow up the significant effect with a **post hoc trend analysis**.

But why should you do a trend analysis to determine the shape of the functional relationship between your independent and dependent variables? Can't you see this relationship by simply graphing the group means? Yes and no. Yes, graphing your sample's means allows you to see the pattern in the data produced by your experiment. No, graphing does not tell you that the pattern you observe represents the true relationship between the variables because your pattern could be due to random error (e.g., if even one mean is thrown off by random error, that one misplaced mean could make a linear relationship look nonlinear). Just as you needed statistics to tell you if the difference between two groups was significant (even though you could easily see whether one mean was higher than the other), you need statistics to know if the pattern you observe in your data (a straight line, a curved line, a combination of a curve and a straight line, etc.) would occur if you repeated the experiment. The statistical test you need to determine whether the pattern in your data reflects a reliable functional relationship is a post hoc trend analysis.

Computing a post hoc trend analysis is easy. You can either follow the simple directions in Appendix F or use a computer program that does the analysis for you. Although you might be tempted to forget about post hoc trend analysis until it comes time to analyze your data, don't make that mistake.

If you do not think about post hoc trend analysis when designing your experiment, you will probably be unable to do a valid post hoc trend analysis on your data. Therefore, if you think that you might want to know about the functional relationship between the variables in your experiment, you should keep three facts in mind *before* conducting that experiment (see Box 11.2).

First, to do a post hoc trend analysis, you should have selected levels of your independent variable that increase proportionally. For example, if you were using three levels of a drug, you would not use 5 mg, 6 mg, and 200 mg. Instead, you might use 10 mg, 20 mg, and 30 mg, or 10 mg, 100 mg, and 1000 mg.

Second, you must have at least an interval scale measure of your dependent variable. Your map of the functional relationship can't be accurate unless your measure of the dependent variable is to scale. If you tried to find the relationship between the loudness of the music playing on participants'

BOX 11.2 Requirements for Conducting a Valid Post Hoc Trend Analysis

1. Your independent variable must have a statistically significant effect.
2. Your independent variable must be quantitative, and the levels used in the experiment should vary from one another by some constant proportion.
3. Your dependent variable must yield interval or ratio-scale data so that your map of the functional relationship will be to scale.
4. The number of trends you can look for is one less than the number of levels of your independent variable.

personal stereos and distance walked, you would have to measure distance by number of meters walked rather than by blocks walked (unless all your blocks are the same length). In short, you can't do a trend analysis if you have ordinal or nominal data.

Third, the more levels of the independent variable you have, the more trends you can look for. Specifically, the number of trends you can examine is one less than the number of levels you have. If you had only two levels, you can test only for straight lines (linear component). If you have three groups, you can test for straight lines (linear component), and for a U-shaped curve (quadratic component). With four levels, you can test for straight lines, U-shaped curves, and double U-shaped lines (cubic component). Thus, if you are expecting a double U-shaped curve, you must use at least four levels of the independent variable.

CONCLUDING REMARKS

By using a multiple-group experiment rather than a simple experiment, you can ask more refined questions. For example, you can go beyond asking, "Is there an effect?" to asking "What is the nature of the functional relationship?"

By using a multiple-group experiment rather than a simple experiment, you can get more valid answers to your questions. For example, by using appropriate control groups, you can learn not only that the treatment manipulation worked but also why it worked.

Although adding more levels of the treatment is a powerful way to expand the simple experiment, an even more powerful way to expand the simple experiment is to add independent variables. As you will see in the next chapter, adding independent variables not only increases construct and external validity but also opens up a whole new arena of research questions.

SUMMARY

1. The multiple-group experiment is more sensitive to nonlinear relationships than the simple experiment. Consequently, it is more likely to obtain significant treatment effects and to accurately map the functional relationship between your independent and dependent variables.
2. Knowing the functional relationship allows more accurate predictions about the effects of unexplored levels of the independent variable.
3. To use the multiple-group experiment to discover the functional relationship, you should select your levels of the independent variable carefully, and your dependent measure must provide at least interval scale data.
4. Multiple-group experiments may have more construct validity than a simple experiment because they can have multiple control groups and multiple treatment groups.
5. To analyze a multiple-group experiment, you first have to conduct an analysis of variance (ANOVA). An ANOVA will produce an *F* ratio.
6. An *F* ratio is a ratio of between-groups variance to within-groups variance.
7. Random error will make different treatment groups differ from each other. If the treatment has an effect, the treatment will also cause the groups to differ from each other. In other words, between-groups variance is due to random error and may also be due to treatment effects. Because it may be affected by treatment effects, between-groups variance is often called treatment variance.

8. Scores within a treatment group differ from each other for only one reason: random error. That is, the treatment cannot be responsible for variability within each treatment group. Therefore, within-groups variance is an estimate of the degree to which random error affects the data. Consequently, another term for within-groups variance is error variance.
9. The *F* test is designed to see whether the difference between the group means is greater than would be expected by chance. It involves dividing the between-groups variance (an estimate of random error plus possible treatment effects) by the within-groups variance (an estimate of random error). If the *F* is 1 or less, there is no evidence that the treatment has had an effect. If the *F* is larger than 1, you need to look in an *F* table (under the right degrees of freedom) to see whether the *F* is significant.
10. The first degrees of freedom (between groups/treatment) equals the number of groups minus one, abbreviated $G-1$. The second degrees of freedom (within groups/error) equals the number of participants minus the number of groups, abbreviated $N-G$. Thus, if you had 5 groups and 40 participants, you would look at the *F* table under 4 (5–1) and 35 (40–5) degrees of freedom.
11. You are most likely to get a significant *F* if between-group variability is large (your groups differ from each other) and within-groups variability is small.
12. If you get a significant *F*, you know that the groups are not all the same. If you have more than two groups, you have to find out which groups differ. To find out which groups are different, do not just look at the means to see which differences are biggest. Instead, do post hoc tests to find out which groups are reliably different.
13. The following table summarizes the mathematics of an ANOVA table.

SOURCE OF VARIANCE (*SV*)	SUM OF SQUARES (*SS*)	DEGREES OF FREEDOM (*DF*)	MEAN SQUARE (*MS*)	*F*
Treatment (*T*)	*SST*	Levels of *T*–1	*SST/df T*	*MST/MSE*
Error (*E*)	*SSE*	Participants–Groups	*SSE/df E*	
Total	*SST*+*SSE*	Participants–1		

KEY TERMS

functional relationship (*p. 387*)
linear relationship (*p. 388*)
confounding variables (*p. 394*)
hypothesis-guessing (*p. 396*)
empty control group (*p. 396*)
variability between group means (*p. 400*)
within-groups variance (*p. 402*)
error variance (*p. 402*)
treatment variance (*p. 404*)
analysis of variance (ANOVA) (*p. 404*)
F ratio (*p. 405*)
eta squared (η^2) (*p. 409*)
post hoc tests (*p. 409*)
post hoc trend analysis (*p. 411*)

EXERCISES

1. A researcher randomly assigns each member of a statistics class to one of two groups. In one group, each student is assigned a tutor who is available to meet with the student 20 minutes before each class. The other group is a control group not assigned a tutor.

Suppose the researcher finds that the tutored group scores significantly better on exams.

a. Can the researcher conclude that the experimental group students learned statistical information from tutoring sessions that enabled them to perform better on the exam? Why or why not?
b. What changes would you recommend in the study?

2. Suppose people living in homes for older adults were randomly assigned to two groups: a no-treatment group and a transcendental meditation (TM) group. Transcendental meditation involves more than sitting with eyes closed. The technique involves both "a meaningless sound selected for its value in facilitating the transcending, or settling-down, process and a specific procedure for using it mentally without effort again to facilitate transcending" (Alexander, Langer, Newman, Chandler, & Davies, 1989, p. 953). The TM group was given instruction in how to perform the technique; then "they met with their instructors half an hour each week to verify that they were meditating correctly and regularly. They were to practice their program 20 minutes twice daily (morning and afternoon) sitting comfortably in their own room with eyes closed and using a timepiece to ensure correct length of practice" (Alexander et al., 1989, p. 953).

   Suppose that the TM group performed significantly better than other groups on a mental health measure.[11]

   a. Could the researcher conclude that it was the transcendental meditation that caused the effect?
   b. What besides the specific aspects of TM could cause the difference between the two groups?
   c. What control groups would you add?
   d. Suppose you added these control groups and then got a significant *F* for the treatment variable? What could you conclude? Why?

3. Assume you want to test the effectiveness of a new kind of therapy. This therapy involves screaming and hugging people in group sessions followed by individual meetings with a therapist. What control group(s) would you use? Why?
4. Assume a researcher is looking at the relationship between caffeine consumption and sense of humor.
   a. How many levels of caffeine should the researcher use? Why?
   b. What levels would you choose? Why?
   c. If a graph of the data suggests a curvilinear relationship, can the researcher assume that the functional relationship between the independent and dependent variables is curvilinear? Why or why not?
   d. Suppose the researcher used the following four levels of caffeine: 0 mg, 20 mg, 25 mg, 26 mg. Can the researcher easily do a trend analysis? Why or why not?
   e. Suppose the researcher ranked participants based on their sense of humor. That is, the person who laughed least got a score of 1, the person who laughed second-least scored a 2, and so on. Can the researcher use these data to do a trend analysis? Why or why not?
   f. If a researcher used four levels of caffeine, how many trends can the researcher look for? What are the treatment's degrees of freedom?
   g. If the researcher used three levels of caffeine and 30 participants, what are the degrees of freedom for the treatment? What are the degrees of freedom for the error term?
   h. Suppose the *F* is 3.34. Referring to the degrees of freedom you obtained in your answer to "g" (above) and to Table 3 (Appendix F), are the results statistically significant? Can the researcher look for linear and quadratic trends?

[11] A modification of this study was actually done. The study included appropriate control groups.

5. A computer analysis reports that $F(6, 23) = 2.54$. The analysis is telling you that the F ratio was 2.54, and the degrees of freedom for the top part of the F ratio = 6 and the degrees of freedom for the bottom part = 23.
   a. How many groups did the researcher use?
   b. How many participants were in the experiment?
   c. Is this result statistically significant at the .05 level? (Refer to Table 3 of Appendix F.)
6. A friend gives you the following Fs and significance levels. On what basis would you want these Fs (or significance levels) rechecked?
   a. F(2, 63) = .10, not significant
   b. F(3, 85) = –1.70, not significant
   c. F(1, 120) = 52.8, not significant
   d. F(5, 70) = 1.00, significant
7. Complete the following table.

SOURCE OF VARIANCE (*SV*)	SUM OF SQUARES (*SS*)	DEGREES OF FREEDOM (*df*)	MEAN SQUARE (*MS*)	*F*
Treatment (*T*) 3 levels of treatment	180	—	—	—
Error (*E*), also known as within-groups variance	80	8	—	

8. Complete the following table.

SOURCE OF VARIANCE (*SV*)	SUM OF SQUARES (*SS*)	DEGREES OF FREEDOM (*df*)	MEAN SQUARE (*MS*)	*F*
Treatment (*T*) (between groups variance)	50	5	—	—
Error (*E*), (within-groups variance)	100	—	—	—
Total	—	30		—

9. A study compares the effect of having a snack, taking a 10-minute walk, or getting no treatment on energy levels. Sixty participants are randomly assigned to a condition and then asked to rate their energy level on a 0 (not at all energetic) to 10 (very energetic) scale. The mean for the "do nothing" group is 6.0, for having a snack 7.0, and for walking 7.8. The F ratio is 6.27.
   a. Graph the means.
   b. Are the results statistically significant?
   c. If so, what conclusions can you draw? Why?
   d. What additional analyses would you do? Why?
   e. How would you extend this study?

WEB RESOURCES

1. Go to the Chapter 11 section of the book's student website and
   a. Look over the concept map of the key terms.
   b. Test yourself on the key terms.
   c. Take the Chapter 11 Practice Quiz.
   d. Download the Chapter 11 tutorial.
2. Do an analysis of variance using a statistical calculator by going to the "Statistical Calculator" link.
3. If you want to write your method section, use the "Tips on Writing a Method Section" link.
4. If you want to write up the results of a one-factor, between-participants experiment, click on the "Tips for Writing Results" link.

CHAPTER 12

Expanding the Experiment

Factorial Designs

I'm an earth sign, she was a water sign—together we made mud.
—Woody Allen

The pure and simple truth is rarely pure and never simple.
—Oscar Wilde

CHAPTER OVERVIEW

To understand the relationship between the design we will discuss in this chapter and the experimental designs we discussed in previous chapters, let's look at three ways you might partially replicate Langer, Blank, and Chanowitz's (1978) classic experiment. In that experiment, research assistants tried to cut in front of participants who were in line to use a copier. Participants were randomly assigned to receive one of several requests.

If you were to replicate that study as a simple experiment, participants would be randomly assigned to one of two requests. For example, if you chose to vary quality of excuse, half your participants might be asked, "Can I cut in front of you?" (no excuse condition), whereas the other half might be asked, "Can I cut in front of you because I want to make a copy?" (nonsensical excuse condition). In the following table, we have diagrammed the design and results of such a simple experiment.

Proportion of Participants Who Agreed to Let the Researcher Cut in Front of Them to Use the Copier as Function of Researcher's Excuse

TYPE OF EXCUSE	
Group 1	**Group 2**
No excuse	Senseless excuse ("I need to make copies")
.60	.93

In Chapter 10, we showed how the simple experiment's logic makes it internally valid. However, we also pointed out that the simple experiment is limited because it can study only two levels of a single independent variable. For example, with a single simple experiment, you cannot compare three different excuse conditions (e.g., no excuse, senseless excuse, and reasonable excuse), three levels of temperature (e.g., cold, medium, hot), or three types of music (e.g., classical, rock, and rap).

In Chapter 11, we showed how to extend the simple experiment's logic to experiments that study three or more levels of a single independent variable. By randomly assigning participants to three or more levels of the

treatment, you can look at the effects of varying three levels of excuses, temperature, music, or any other variable. For example, by adding a level to the excuse experiment diagrammed earlier, you can expand it into the three-group experiment diagrammed here:

Proportion of Participants Who Agreed to Let the Researcher Cut in Front of Them to Use the Copier as Function of Researcher's Excuse

TYPE OF EXCUSE		
Group 1	**Group 2**	**Group 3**
No excuse	Senseless excuse ("I need to make copies")	Reasonable excuse ("because I'm in a rush")
.60	.93	.94

As we discussed in Chapter 11, experiments that manipulate three or more levels of a factor can have impressive internal, external, and construct validity.

In this chapter, as in Chapter 11, we show how to extend the basic logic of the simple experiment. However, instead of showing you how to stretch the simple experiment by adding more levels of a factor, we show you how to expand the simple experiment by adding more factors. For example, rather than learning how to expand a simple experiment on excuses to include more than two types of excuses, you will learn how to add another factor, such as size of request, so you can study the effects of both excuses and request size in the same experiment (see the following diagram).

Proportion of Participants Who Agreed to Let Researcher Cut in Front of Them to Use the Copier as Function of Excuse.

	TYPE OF EXCUSE		
Size of Request	**No Excuse**	**Senseless Excuse ("I need to make copies")**	**Reasonable Excuse ("because I'm in a rush")**
Small ("I have 5 pages")	.60	.93	.94
Large ("I have 20 pages")	.24	.24	.42

Note: Data are from Langer, Blank, and Chanowitz (1978).

In technical terms, you will learn about **factorial experiments**: experiments that study the effects of two or more independent variables (***factors***) in a single experiment. Specifically, you will learn (a) why you should want to study the effects of two independent variables in a single experiment, (b) how to design such experiments, and (c) how to analyze the results of such experiments.

THE 2 × 2 FACTORIAL EXPERIMENT

To understand why and how to design factorial experiments, we will focus on the simplest factorial experiment: the 2 × 2 ("2 by 2") between-subjects factorial experiment. Before discussing why you would want to do a 2 × 2 factorial experiment, let's be clear about how the 2 × 2 is similar to and different from other factorial experiments.

Although all factorial experiments must include at least two levels of two factors, factorial experiments can differ in (a) how many levels of each factor they have and (b) how many factors they have. To let people know how many levels each factor has, researchers use terminology similar to what builders use. When a builder refers to a "2 by 4," the builder means a board for which the first dimension (thickness) is 2 inches and the second dimension (width) is 4 inches. Similarly, when a researcher refers to a "2 by 4," the researcher means that the first experimental factor has 2 levels and the second experimental factor has 4 levels. Thus, the Langer, Blank, and Chanowitz (1978) we described earlier was a 3 (Excuse type: no excuse, poor excuse, or reasonable excuse) × 2 (Request size: small or large).

In a 2 × 2 factorial experiment, there are two independent variables and both have two levels. For example, suppose we had a 2 (Excuse type: no excuse or reasonable excuse) × 2 (Size of request: small or large) experiment. The "×"—pronounced as "by"—indicates that the first variable is crossed (combined) with the second factor. That is, rather than conditions consisting of only a single manipulation (e.g., no excuse or reasonable excuse), each condition will consist of a manipulation of the first factor (e.g., no excuse or reasonable excuse) combined with a manipulation of the second factor (e.g., small request or large request). Thus, in a 2 (Excuse: none, reasonable) × 2 (Size of request: small, large) factorial, crossing 2 levels of 2 different independent variables would result in 4 (2 × 2) different treatment conditions: (1) a no excuse, small request condition; (2) a no excuse, large request condition; (3) a reasonable excuse, small request condition; and (4) a reasonable excuse, large request condition (see the next table).[1]

Size of Request	No Excuse	Reasonable Excuse ("because I'm in a rush")
Small ("I have 5 pages")	.60	.94
Large ("I have 20 pages")	.24	.42

In the 2 × 2 *between-subjects* factorial experiment, each participant is randomly assigned to experience one—and only one—of the four treatment combinations. Thus, in the example diagrammed previously, you would have four groups: (1) a no excuse, small request group; (2) a no excuse, large request group; (3) a reasonable excuse, small request group; and (4) a reasonable excuse, large request group.

[1] If we had 3 levels of excuse instead of just 2, we would have a 3 × 2 design instead of a 2 × 2. With a 3 × 2, we would have 6 (3 × 2) different conditions. If we had three, 2-level factors (excuse type, request size, and gender of experimenter), we would have a 2 × 2 × 2 design. With a 2 × 2 × 2 experiment, we would have 8 (2 × 2 × 2) experimental conditions.

To better understand how a 2 × 2 between-subjects factorial experiment works, let's turn to an actual 2 × 2 experiment: Pronin and Wegner's (2007) experiment on manic thinking. In that experiment, the researchers were interested in seeing whether getting participants' thoughts to race would boost participants' moods—and whether this boost would occur even when people were thinking negative thoughts. To manipulate what participants thought, Pronin and Wegner had participants read aloud 60 statements that were either uplifting or depressing. To control how fast participants were thinking, Pronin and Wegner made participants read those statements either very quickly or very slowly. Participants randomly assigned to the uplifting statements groups read a neutral statement—"Today is no better or worse than another day"—and then read statements that became increasingly positive. For example, the second statement participants in the uplifting statements group read was "I do feel pretty good today, though," whereas the last statement they read was "Wow! I feel great!" Participants randomly assigned to the depressing statements read the same neutral statement as the uplifting statements group ("Today is no better or worse than any other day") but then read statements that became increasingly negative. For example, the second statement they read was "However, I feel a little low today," whereas the last statement they read was "I want to go to sleep and never wake up."

Half of the participants in the uplifting statements condition were randomly assigned to read the statements quickly (about twice as fast as students would normally read those statements) whereas the other half were to read the statements slowly (about half as fast as students would normally read those statements). Similarly, half the participants in the depressing statements condition were randomly assigned to read the statements quickly, whereas the other half were randomly assigned to read the statements slowly. Both fast and slow condition participants read statements aloud from a PowerPoint® presentation: The only difference was that the PowerPoint® presentation went nearly four times as fast in the fast condition as in the slow condition. After the participants read the statements, they filled out several scales, one of which was a mood scale. Thus, if you were to repeat the Pronin and Wegner (2007) 2 (Statement type: negative or positive) × 2 (Statement speed: slow or fast), you would randomly assign participants so that one-fourth of your participants were in each of the four conditions diagrammed in the following table:

GROUP 1	GROUP 2
Negative statements	Negative statements
Slow presentation	Fast presentation
GROUP 3	**GROUP 4**
Positive statements	Positive statements
Slow presentation	Fast presentation

Each Column and Each Row of the 2 × 2 Factorial Is Like a Simple Experiment

You could view each *row* of the 2 × 2 factorial as a simple experiment. With that perspective, you would see the 2 × 2 factorial experiment as two simple experiments, both of which looked at whether participants are in better moods when statements are presented quickly than when statements are presented slowly. That is, as you can see from the following table, both experiments compare slow presentation to fast presentation.

SIMPLE EXPERIMENT 1 (Effect of slow vs. fast presentation for negative statements)	GROUP 1	GROUP 2
	Negative statements *Slow presentation*	*Negative statements* *Fast presentation*

SIMPLE EXPERIMENT 2 (Effect of slow vs. fast presentation for positive statements)	GROUP 3	GROUP 4
	POSITIVE STATEMENTS Slow presentation	POSITIVE STATEMENTS Fast presentation

You could also view each *column* of the 2 × 2 factorial as a simple experiment. With that perspective, you would see the 2 × 2 factorial experiment as two different simple experiments, both of which looked at whether participants are in a better mood after reading positive statements than after reading negative statements (see the following table).

SIMPLE EXPERIMENT 3 (EFFECT OF STATEMENT TYPE [NEGATIVE VS. POSITIVE] IN THE SLOW CONDITIONS)	SIMPLE EXPERIMENT 4 (EFFECT OF STATEMENT TYPE [NEGATIVE VS. POSITIVE] IN THE FAST CONDITIONS)
GROUP 1	GROUP 2
Negative statements Slow presentation	*Negative statements* Fast presentation
GROUP 3	GROUP 4
Positive statements Slow presentation	*Positive statements* Fast presentation

If you looked at both the rows and the columns, you would see that the factorial experiment contains four simple experiments (see the following table).

	COLUMN CONTAINING SIMPLE EXPERIMENT 3	COLUMN CONTAINING SIMPLE EXPERIMENT 4
	(Effect of negative vs. positive statements in slow presentation conditions)	(Effect of *negative vs. positive statements* in fast presentation conditions)
ROW CONTAINING SIMPLE EXPERIMENT 1	GROUP 1	GROUP 2
(Effect of **slow vs. fast presentation** for negative statement participants)	Negative statements **Slow presentation**	*Negative statements* **Fast presentation**
ROW CONTAINING SIMPLE EXPERIMENT 2	GROUP 3	GROUP 4
(Effect of **slow vs. fast presentation** for positive statement participants)	Positive statements **Slow presentation**	*Positive statements* **Fast presentation**

How One Experiment Can Do More Than Two

To illustrate how similar each row of a 2 × 2 is to a simple experiment, suppose you had done a simple experiment involving only the two groups listed in the first row of the 2 × 2 (the negative statements/**slow presentation** group vs. the negative statements/**fast presentation** group). In that case, you would see the effect of, as the authors put it, "thinking slowly" vs. "thinking fast," for participants who read only negative statements. In the same way, if you did the 2 × 2 experiment diagrammed above and compared only the two groups in the first row of the 2 × 2 (the negative statements/**slow** presentation group vs. the negative statements/**fast** presentation group), you would get the **simple main effect** of presentation speed for participants who read only negative statements.

The 2 × 2 Yields Four Simple Main Effects

Because the 2 × 2 contains four simple experiments, if we used certain statistical techniques, we could use the 2 × 2 to find four simple main effects:

1. the simple main effect for **speed** in the negative statements conditions (by looking at the first row and comparing the **slow presentation,** negative statements group with the **fast presentation,** negative statements group)
2. the simple main effect for **speed** in the positive statements conditions (by looking at the second row and comparing the **slow presentation,** positive statements group with the **fast presentation,** positive statements group)
3. the simple main effect for negative vs. positive statements in the slow presentation conditions (by looking at the first column and comparing the negative statements, slow presentation group with the positive statements, slow presentation group)

4. the simple main effect for *negative vs. positive statements* in the fast presentation conditions (by looking at the second column and comparing the *negative statements*, fast presentation group with the *positive statements,* fast presentation group)

The simplest way to *estimate* these simple main effects is to subtract the relevant group means from each other. To illustrate, suppose the cell means for our four groups were as follows:

GROUP 1	GROUP 2
Negative statements Slow presentation	*Negative statements* *Fast presentation*
4	*6*

GROUP 3	GROUP 4
Positive statements Slow presentation	*Positive statements* *Fast presentation*
12	*14*

With these means, we could estimate four simple main effects: two speed (slow vs. fast) simple main effects (by looking at the two rows) and two statement type (positive vs. negative) simple main effects (by looking at two columns). Let's first look for the two speed simple main effects by comparing the groups that differ in terms of speed but are the same in terms of whether they read positive or negative statements:

1. The simple main effect for **speed** in the negative statements conditions = 2 (*6* – 4; see the first row).
2. The simple main effect for **speed** in the positive statements conditions = 2 (*14* – 12; see the second row).

Now, let's look for the two statement type simple main effects by comparing the groups that are different in terms of statement type but are the same in terms of speed:

3. The simple main effect for statement type (positive vs. negative) in the slow statements conditions = 8 (12 – 4); see the first column).
4. The simple main effect for *statement type* in the fast statements conditions = 8 (*14* – *6*; see the second column).

The following table displays the group means and *estimates* of our four simple main effects.

	SLOW SPEED	FAST SPEED	SPEED SIMPLE MAIN EFFECTS
Negative statements	4 (Group 1)	*6 (Group 2)*	+2 (6 – 4)
Positive statements	12 (Group 3)	*14 (Group 4)*	+2 (*14* – 12)
Statement type simple main effects	+8 (12 – 4)	+8 (*14* – *6*)	

The 2 × 2 Yields Two Pairs of Simple Main Effects

We have shown you that the 2 × 2 can yield four simple main effects. However, the strength of the 2 × 2 is not that it produces four separate main effects. Instead, its strength is that it produces two *pairs* of simple main effects: (1) a pair of simple main effects relating to the first independent variable (e.g., two speed simple main effects) and (2) a pair of simple main effects relating to the second independent variable (e.g., two type of statement [uplifting vs. depressing] simple main effects). To capitalize on the two pairs of simple main effects that the 2 × 2 produces, researchers' analyses focus on two things:

1. combining (*averaging*) each treatment's pair of simple main effects to estimate each treatment's overall, average effect
2. contrasting (*subtracting*) each treatment's pair of simple main effects to determine whether the treatment has one effect on one group of participants but a different effect on a different group of participants

Averaging a Treatment's Simple Main Effects Lets You Estimate the Overall Main Effect: The Average Effect of Varying a Factor

To combine a treatment's simple main effects, you average them. The *average* of a treatment's two simple main effects allows you to estimate the treatment's ***overall main effect***: the average effect of varying that treatment.

In the 2 (Speed of thought: slow or fast) × 2 (Type of thought: positive or negative), the researcher would average the two speed simple main effects to get an estimate of the overall main effect for speed. To illustrate, suppose the simple main effect of presentation speed was +2 in the negative statements condition (the fast presentation, negative statements participants scored 2 points higher on the mood scale than the slow presentation, negative statements participants). Furthermore, suppose that the simple main effect of presentation speed was +4 in the positive statement conditions (the fast presentation, positive statements participants scored 4 points higher on the mood scale than the slow presentation, positive statements participants). In that case, the estimate for the overall main effect of presentation speed would be 3 (because the average of 2 and 4 is 3).

Similarly, to estimate the overall main effect for (negative vs. positive) statement type, the researcher would average the two statement type simple main effects. If the overall statement type effect was statistically significant, it would mean that, on the average, participants who read negative statements were in a different mood than the participants who read positive statements.

One reason researchers emphasize overall main effects is convenience. It is easier to talk about one overall main effect than about two simple main effects.

However, a more important reason for averaging the two simple main effects into an overall main effect is that it allows us to make more general statements about that variable's effects. Consider the advantage of averaging the two simple speed main effects. Because we combined two simple main effects, we are not confined to saying that speeding up thoughts improves mood if you are already thinking positive thoughts. Instead, we can say that, on the average, across conditions that varied from participants thinking

negative thoughts to participants thinking positive thoughts, participants who thought faster were in better moods.

Subtracting a Treatment's Simple Main Effects Lets You Estimate the Interaction

But what if the simple main effect for speed of thought is different in the negative thought condition than in the positive thought condition? Then:

a. You should *not* make a general statement about the effects of thought speed without mentioning that the effect of speeding up thought changes depending on whether the person is thinking negative thoughts or positive thoughts.
b. You should be happy that you can compare thought speed's simple main effects with each other because that comparison lets you know that the effect of speeding up thought depends on whether the person is thinking negative thoughts or positive thoughts.

By comparing the two simple main effects of speed (the speed simple main effect for the negative statements condition and the speed simple main effect for the positive statements condition), you would be able to tell whether the effect of speeding up thoughts *depended on* whether participants are thinking positive or negative thoughts. If, for example, you found that that speeding up thoughts had a negative effect in the negative statements condition, but had a positive effect in the positive statements condition, you could say that the effect of speeding up thoughts depends on the type of statements participants read.

If the simple main effects of speed differ *depending* on the type of statement (positive or negative), there is an **interaction** between speed and statement type (see Table 12.1). If, on the other hand, speed's simple main effects do not differ from each other (speed has the same effect in the negative statements condition as it has in the positive statements condition), you do not have an interaction. If you do not have an interaction, the effect of combining those variables is what you would expect from adding up their individual effects.

Why You Want to Look for Interactions: The Importance of Moderating Variables

Interactions are important and common (see Table 12.2). Treatments will tend to have one effect on one group but another effect on another group. For example, eating grapefruit is good for most people, but not for people who are taking certain kinds of medications. For those people, eating grapefruit may kill them. For them, the positive main effect for eating grapefruit is unimportant relative to the dangerous grapefruit × drug interaction.[2]

Interactions do not have to be dangerous. The only requirement for an interaction is that the effect of combining treatments is different from the sum of their individual effects. For example, there is an interesting interaction involving caffeine and nicotine, both of which are stimulants. Consuming caffeine increases physiological arousal—unless people have nicotine in their

[2] A popular and effective allergy medicine was taken off the market because of this deadly interaction.

TABLE **12.1**
Simple Main Effects, Overall Main Effects, and Interactions

SIMPLE MAIN EFFECTS	
Definition	The effects of one independent variable at a specific level of a second independent variable. The simple main effect could have been obtained merely by doing a simple experiment.
How to Estimate	Compare the mean for one group with the mean for a second group (for instance, comparing the average for the *slow* thoughts, negative thoughts group to the average for the *fast* thoughts, negative thoughts group).
Question Addressed	What is the effect of the thought speed in the negative statements condition?
OVERALL MAIN EFFECT	
Definition	The average effect of a treatment.
How to Estimate	Average a treatment's simple main effects. If the average of the two simple main effects is significantly different from zero, there is an overall main effect.
Question Addressed	What is the average effect of speeding up thoughts in this study?
INTERACTION	
Definition	The effect of a treatment is different, depending on the level of a second independent variable. That is, the effect of a variable is uneven across conditions.
How to Estimate	Look at the *differences* between a treatment's simple main effects. If the treatment's simple main effects are the same, there is no interaction. If, however, the treatment's two simple main effects differ significantly, there is an interaction.
Question Addressed	Does speeding up thoughts have a different effect on those who read negative statements than it has on those who read positive statements?

system. For people who have a lot of nicotine in their system, caffeine actually reduces physiological arousal: The person who has smoked several cigarettes can wind down by drinking a caffeinated cola.[3]

Interactions do not have to involve reversing the treatment's original effect. To have an interaction, all that is required is that the effect of combining the treatments has an effect that is different from the sum of their individual effects. Thus, if two drugs each have a mild positive effect but taking both drugs together has an enormously positive effect, you have an interaction. Likewise, if one drug has a mild positive effect and another drug has no measurable effect, but taking both drugs together has an enormous effect, you have an interaction.

If neither drug has a measurable effect by itself but taking both drugs together has a strong effect, you have an interaction. If either drug by itself has a moderate positive effect but taking both drugs together still has no more than a moderate positive effect, you have an interaction. If either drug by itself has a moderate positive effect, but taking both drugs together has no effect, you have an interaction. In short, whether the relationship between

[3] We are indebted to an anonymous reviewer for this example.

TABLE **12.2**
Ways of Thinking About Interactions

VIEWPOINT	HOW VIEWPOINT RELATES TO INTERACTIONS
Chemical Reactions	Lighting a match, in itself, is not dangerous. Having gasoline around is not, in itself, dangerous. However, the *combination* of lighting a match in the presence of gasoline is explosive. Because the explosive effects of combining gas and lighting a match are different from simply adding their separate, individual effects, gasoline and matches interact.
Personal Relationships	John likes most people. Mary is liked by most people. *But* John dislikes Mary. Based only on their individual tendencies, we would expect John to like Mary. Apparently, however, like gasoline and matches, the combination of their personalities produces a negative outcome.
Sports	A team is not the sum of its parts. The addition of a player may do more for the team than the player's abilities would suggest—or the addition may help the team much less than would be expected because the addition upsets team "chemistry." In other words, the player's skills and personality may interact with those of the other players on the team. Knowing the interaction between the team and the player—how the two will mesh together—may be almost as important as knowing the player's abilities. Good pitchers get batters out. Poor hitters are easier to get out than good hitters are. *However,* sometimes a poor hitter may have a good pitcher's "number" because the pitcher's strengths match the hitter's strengths. Similarly, some "poor" pitchers are very effective against some of the league's best batters. Managers who can take advantage of these interactions can win more games than would be expected by knowing only the talents of the individual team members.
Prescription Drugs	Drug A may be a good, useful drug. Drug B may also be a good, useful drug. However, taking Drug A and B together may result in harm or death. Increasingly, doctors and pharmacists have to be aware of not only the effects of drugs in isolation but also of their combined effects. Ignorance of these interactions can result in deaths and in malpractice suits.
Making General Statements	Interactions indicate that you cannot talk about the effects of one variable without mentioning that the effect of that variable depends on a second variable. Therefore, if you have an interaction, when discussing a factor's effect, you need to say "but," "except when," "depending on," "only under certain conditions." Indeed, you will often see results sections say that the main effect was "qualified by a _____ interaction" or "the effect of the _____ variable was different depending on the level of (the other) variable."
Visually	If you graph an interaction, the lines will not be parallel. That is, the lines either already cross or if they were extended, they would eventually cross.

(*continued*)

TABLE **12.2**
Ways of Thinking About Interactions (Continued)

VIEWPOINT	HOW VIEWPOINT RELATES TO INTERACTIONS
Mathematically	If you have an interaction, the effect of combining the variables is *not* the same as adding their two effects. Rather, the effect is better captured as the result of multiplying the two effects. That is, when you add 2 to a number, you know the number will increase by 2, regardless of what the number is. However, when you multiply a number by 2, the effect will depend on the other number. When doubling a number, the effect is quite different when the number to be doubled is 4 than when it is 1,000 or than when it is –40. To take another example of the effect of multiplication, consider the multiplicative effects of interest rates on your financial condition. If interest rates go up, that will have a big, positive effect on your financial situation if you have lots of money in the bank; a small, positive effect if you have little money in the bank; and a negative effect on your finances if you owe money to the bank (you will have to pay more interest on your debt).

ments could be characterized as "better apart" (two is less than one), "it takes two" (alone they are nothing), "better together" (two is more than one plus one), or "one is enough" (one plus one only equals one), *as long as the effect of combining treatments is different from the sum of their individual effects, you have an interaction.*

In addition to knowing about drug interactions, most people suspect that the effect of an action depends on (*interacts* with) other factors. For example:

- Most people know that telling someone "congratulations" will have a good effect if she has just been promoted but a bad effect if she has just been fired.
- Most people suspect that, under some conditions, it pays to accuse others of something, but under some conditions, accusing others may backfire.

Research supports the popular notion that some treatments will have one effect on one group of participants, but a different effect on another group. For example, Rucker and Petty (2003) found that, of the two groups of participants who read about an employee who had a *bad* work ethic, the group that learned that the employee had accused his coworkers of having a bad work ethic liked the employee *more* than did the group that did not learn of the employee making such accusations. On the other hand, of the two groups of participants who read about an employee who had a *good* work ethic, the participants who learned that the employee had accused his coworkers of having a bad work ethic liked the employee *less* than did the participants who were not told that the employee had made any accusations. Thus, there was an employee reputation × accusation interaction.

The previous example illustrates that interactions—the effects of a combination of treatments being different from the sum of those variables' individual effects—may involve social variables. Note, however, that interactions can involve any variables—even physical variables such as noise and lighting. For instance, consider the effects of two manipulated variables: (1) noise level and (2) perception of control. If you make a group of participants believe they

have no control over the noise level in the room, increasing the noise level seriously harms performance. But for participants led to believe that they can control the noise level, increasing the noise level does *not* harm performance. Thus, noise level interacts with perceived control (Glass & Singer, 1972).

Because of this interaction between noise level and perceived control, you cannot simply say that noise hurts performance. You have to say that the effect of noise level on performance *depends* on (is moderated by) perceived control. In other words, rather than stating a simple rule about the effects of noise, you have to state a more complex rule. This complex rule puts qualifications on the statement that noise hurts performance. Specifically, the statement that noise hurts performance will be qualified by some phrase such as "depending on," "but only if," or "however, that holds only under certain conditions." In short, as Gernsbacher (2007) puts it, if the rule suggested by a main effect is like the spelling rule "*i* before *e*," the rule describing an interaction is more like "*i* before *e* except after *c*." Note that both in the case of spelling and real life, the rule described by the interaction is not as simple as the main effect, but it is more accurate. Thus, as Stanovich (2007) points out, interactions encourage us to go beyond simplistic "either/or" thinking (e.g., is your performance due to your personality or your environment) to "and" thinking (e.g., how is your performance affected by your personality, the environment, and the interaction between your personality and the environment).

Because the concept of interaction is so important, let's consider one more example. As a general rule, we can say that getting within 12 inches (30 cm) of another person will make that person uncomfortable. Thus, the main effect of getting physically closer to someone is to produce a negative mood. However, what if the person who comes that close is extremely attractive? Then, getting closer may elicit positive feelings. Because the effect of interpersonal distance is moderated by attractiveness, we can say that there is an interaction between distance and attractiveness.

In short, you now know two facts about interactions. First, if there is an interaction involving your treatment, it means that the treatment has one effect under one set of conditions but another effect under another set of conditions. Second, interactions play an important role in real life because in real life, the right answer often depends on the situation.

Interesting Questions in Modern Psychology Are Often Questions About Interactions

As psychology has progressed, psychologists have focused increasingly more attention on interactions. One reason psychologists focus on interactions is that psychologists have already discovered the main effects of many variables. We know how most individual variables act in isolation. Now, it is time to go to the next step—addressing the question, "What is the effect of combining these variables?" Put another way, once we learn what the general effect of a variable is, we want to find out what specific conditions may modify (moderate) this general, overall effect. Consequently, in Chapter 3, we encouraged you to generate research ideas that involved moderating variables. In other words, we encouraged you to do what many psychologists do—focus on interactions rather than main effects.

Another reason psychologists focus on interactions is that interactions are common. Consequently, psychologists now frame general problems and issues

in terms of interactions. Rather than asking, "What is the (main) effect of personality and what is the (main) effect of the situation?" psychologists are now asking, "How do personality and the situation interact?" Asking this question has led to research indicating that some people are more influenced by situational influences than others (Snyder, 1984).

Similarly, rather than looking exclusively at the main effects of heredity and the main effects of environment, many scientists are looking at the interaction between heredity and environment. In other words, rather than asking, "What is the effect of a certain environment?" they are asking, "Are the effects of a certain environment different for some people than for others?"

Looking for these interactions sometimes produces remarkable findings. For example, psychologists have found that certain children may thrive in an environment that would harm children who had inherited a different genetic predisposition (Plomin, 1993). Eventually, such research may lead to new ways of educating parents. For instance, rather than telling parents the one right way to discipline children, parent education may involve teaching parents to identify their child's genetic predispositions and then alter their parenting strategies to fit that predisposition. In short, much of the recent research in psychology has involved asking questions that relate to interactions, such as "Under what conditions do rewards hurt motivation?"

External Validity Questions Are Questions About Interactions

We do not mean to imply that the interest in interactions is an entirely new phenomenon. Anyone interested in external validity is interested in interactions. If you are concerned that a treatment won't work on a certain type of person (women, minorities, retired adults), you are concerned about a treatment × type of person interaction. If you are concerned that a treatment that worked in one setting (a hospital) won't have the same effect in a different setting (a school), you are concerned about a treatment × setting interaction. If you are concerned that a treatment won't have the same effect in another culture, you are concerned about a treatment × culture interaction. If you are concerned that the superiority of one treatment over another will diminish over time, you are concerned about a treatment × time interaction. In summary, determining the external validity of your findings is often a matter of determining whether your treatment interacts with time, setting, culture, or type of participant.

Questions in Applied Psychology Are Often Questions About Interactions

Understandably, applied psychologists have always been interested in interactions. One of the founders of applied psychology, Walter Dill Scott, was fascinated by the fact that some people will like an advertisement that others will hate. Therefore, he investigated personality × type of ad interactions.

Most applied psychologists have shared Scott's interest in determining which treatments work on which type of people. For example, therapists know that a therapeutic approach (behavior therapy, drug therapy) that works well for some patients (e.g., individuals with phobias) may not work as well for others (e.g., individuals who are depressed). In other words, good therapists know about treatment × type of patient interactions.

In conclusion, the applied psychologist is keenly interested in interactions. When clients pay for advice, they do not want the expert to know only about

TABLE **12.3**
Questions Addressed by a 2 × 2 Experiment

EFFECT	QUESTION ADDRESSED
Overall main effect for speed	"On the average, does varying speed have an effect?"
Overall main effect for statement type	"On the average, does varying statement type have an effect?"
Interaction between speed and statement type	"Does the effect of speed *differ depending on* what type of statements (positive vs. negative) participants read?" Put another way, "Does the effect of statement type (positive vs. negative) *differ depending on* whether participants are in the slow vs. fast condition?"

main effects. That is, they do not want the expert to stop at saying, "My recommended course of action works in the average case, and so it may work for you." Instead, clients may quiz the expert about interactions involving the expert's proposed treatment. For example, they may ask, "Are there circumstances in which this treatment might make things worse—and does my case fit those circumstances?" To answer this question—that is, to know when a treatment will be helpful and when it will be harmful—the expert must know about the interactions involving that treatment.

Examples of Questions You Can Answer Using the 2 × 2 Factorial Experiment

Now that you have a general understanding of main effects and interactions, let's apply this knowledge to a specific experiment. If you were to replicate Pronin and Wegner's (2007) 2 (Statement type: positive statements vs. negative statements) × 2 (Speed: slow vs. fast) experiment we described earlier, you would look for three different kinds of effects (see Table 12.3).

First, you could look at the main effect of statement type: statement type's *average* effect. You could estimate the overall main effect for statement type by *averaging* the two statement type simple main effects. For example, if, on the average, positive statement participants were in a better mood than participants who read negative statements, you would have a statement type main effect.

Second, you could look at the main effect of speed: speed's average effect. You could estimate the overall main effect for speed by *averaging* the two speed simple main effects. For example, if, on the average, participants who were in the fast thought groups were in a better mood than participants in the slow thought conditions, you would have a speed main effect.

Third, you could look at the interaction between speed and statement type: the extent to which speed's effect *differs* depending on what type of statement participants read. You could probably imagine at least four scenarios that would lead to an interaction:

1. If speeding up thoughts *intensifies* the effect of the statements, speeding up thoughts would, in the negative statement groups, make participants' moods *more negative* but, in the positive statement groups, make participants' moods *more positive*.

2. If speeding up thoughts *weakens* the effects of the statements (perhaps because the participants in the fast condition don't have time to process the statements as much as participants in the slow condition), speeding up thoughts would, for negative thought groups, make participants' moods *less negative* but, for the positive thought groups, make participants' moods *less positive*.
3. If the only way to create manic thinking is to give participants *both* fast thoughts and positive thoughts, speeding up thoughts might only change mood in the positive thoughts condition. Put another way, positive thoughts might only improve mood in the fast condition.
4. If speeding up thoughts and thinking positive thoughts both use the same mechanism to boost mood (e.g., both distract participants from negative thoughts), the group getting *both* positive statements and fast presentation might not do better than the groups getting *either* positive statements or fast presentation.

As we have discussed, if there is an interaction, the effect of combining fast presentation with negative thoughts might be less, more, or even the reverse of what you would expect from knowing only the individual effects of speed and thought type. To begin to estimate the size and type of your interaction, you can *subtract* the two speed simple main effects from each other to get the *difference* between them.

If there is no difference between the two speed simple main effects, there is no interaction: Speed's simple main effects are both the same, and the effect of speed does not depend on type of statement type. Without an interaction, if speed boosts mood by 2 points in the positive statement conditions, it also boosts mood by 2 points in the negative statements conditions.

To review, a significant main effect for statement type would mean that, on the average, varying statement type had an effect on mood. A significant main effect for speed would mean that, on the average, varying speed had an effect on mood. Finally, a significant interaction would mean that the combination of statement type and speed produces an effect that is different (more, less, or opposite) from what you would expect from knowing only statement type's and speed's separate effects.

To illustrate that an interaction indicates that the combination of factors has an effect that is different from the sum of the factor's individual effects, imagine the following situation. Suppose the average effect of positive statements was to boost mood by 2 points and the average effect of fast presentation was also to boost mood by 2 points. If we asked you to guess how much better mood the participants who had the advantages of both receiving positive statements as well as a fast presentation speed (the positive statements/fast presentation group) were in relative to the participants who had neither of these advantages (the negative statements/slow presentation participants), you might, after adding up the effects of positive statements (+2) and fast statements (+2), say "4." In other words, you would guess that, in this case, $2 + 2 = 4$. If there is no interaction, your guess would be right.

But if there is an interaction, your guess would be wrong: The positive statements/fast presentation participants would *not* have a mood that averaged 4 points higher than the mean for the negative statement/slow presentation

participants. If the interaction magnified the effects of the two factors, the positive statements/fast presentation participants might, on the average, score 6 points higher on the mood scale than the negative statements/slow presentation participants.

If, on the other hand, the interaction reversed the effect of the two factors, the positive statements/fast presentation participants might, on the average, score 2 points *lower* than the negative statements/slow presentation participants. If the interaction was the result of one factor neutralizing the effect of another, the positive statements/fast presentation participants might, on the average, score no (0) points higher on the mood scale than the negative statements/slow presentation participants. In short, if you had a statement type × speed interaction, you couldn't predict the mood of the positive statements/fast presentation group merely by adding the statement type effects to the speed effects.

As you can imagine, significant interactions force scientists to answer such questions as, "Does working in groups cause people to loaf?" by saying, "Yes, but it depends on . . ." or "It's a little more complicated than that." Psychologists do not give these kinds of responses to make the world seem more complicated than it is.

On the contrary, psychologists would love to give simple answers. Like all scientists, psychologists prefer parsimonious explanations (simple, elegant explanations that involve few principles) to more complex explanations. Therefore, psychologists would love to report main effects that are not qualified by interactions. Psychologists would like to say that speeding up people's thoughts always increases mood. However, if interactions occur, scientists have the obligation to report them—and in the real world, interactions abound. Only the person who says "Give me a match; I want to see if my gas tank is empty" is unaware of the pervasiveness of interactions. Most of us realize that when variables combine, the effects are different from what you would expect from knowing only their individual, independent effects.

Because we live in a world where we are exposed to more than one variable at a time and because the variables we are exposed to often interact, you may be compelled to do an experiment that captures some of this complexity. But how would you describe the results from such a factorial experiment?

POTENTIAL RESULTS OF A 2 × 2 FACTORIAL EXPERIMENT

You would describe the results of a 2 × 2 factorial experiment in terms of (1) whether you had a main effect for your first independent variable, (2) whether you had a main effect for your second independent variable, and (3) whether you had an interaction. As you can see from Table 12.4, getting a main effect for your first independent variable does not mean that you will be more likely to get a main effect for your second independent variable or that you will be more likely to get an interaction. Instead, like the outcomes of three separate coin flips, the outcomes for the three different effects are independent. Thus, as you can see from Table 12.4 (and as is also true with three separate coin flips), there are eight basic patterns of results you could obtain.

If you did a study, how would you know which of these patterns of results you obtained? At some point, you would need to do a statistical analysis, such as an analysis of variance (ANOVA). Without such a statistical analysis, the patterns you observed in your data might be due to random error rather than

TABLE **12.4**
Eight Potential Outcomes of a 2 × 2 Factorial Experiment

1. A Main Effect for Variable 1	No Main Effect for Variable 2	No Interaction
2. No Main Effect for Variable 1	A Main Effect for Variable 2	No Interaction
3. A Main Effect for Variable 1	A Main Effect for Variable 2	No Interaction
4. A Main Effect for Variable 1	A Main Effect for Variable 2	An Interaction
5. No Main Effect for Variable 1	No Main Effect for Variable 2	An Interaction
6. A Main Effect for Variable 1	No Main Effect for Variable 2	An Interaction
7. No Main Effect for Variable 1	A Main Effect for Variable 2	An Interaction
8. No Main Effect for Variable 1	No Main Effect for Variable 2	No Interaction

Note that having (or not having) a main effect has no effect on whether you will have an interaction.

to statistically reliable treatment effects. Either before or after doing such an analysis, however, you would probably like to see what patterns exist in your data. Therefore, you might calculate the mean response for each group and then make a table of those means. In the next section, we will show you how those tables of means can help you make sense of your results.

One Main Effect and No Interaction

Let's start by supposing you replicate the Pronin and Wegner (2007) experiment we discussed earlier. Using a 2 (positive statements vs. negative statements) × 2 (slow speed vs. fast speed) factorial experiment, suppose you found results like the ones displayed in Table 12.5. To understand your results, you might start looking at the experiment as though it were four separate simple experiments. Thus, if you look only at the first row, it is just like you are looking at the effects of speed in a simple experiment in which all participants read negative statements.

As you can see from the first row of Table 12.5, the slow speed/negative statements group was in the same mood (6) as the fast speed/negative statements group. Thus, varying speed had no noticeable effect in the negative statements condition.

To find out what happened in the positive statements groups, look at the second row. Note that looking at the second row is just like looking at a simple experiment that varied speed (while making all the participants read positive statements). As you can see by the fact that both the slow presentation and the fast presentation scored the same on the mood scale (8), varying speed had no noticeable effect in the positive statement condition.

Averaging the effect of speed over both the negative statements and the positive statements conditions, you find that speed's average (overall) effect was zero. Put another way, the slow speed groups' scores, on the average, were the same as the high speed groups'. Thus, there was no overall main effect for the speed manipulation.

Looking at the columns tells you about the effect of varying whether statements were negative or positive. For example, looking at the first column is like looking at a simple experiment that varied statement type (while having all participants read the statements slowly). As you can see, the positive

TABLE **12.5**
Main Effect for Statement Type, No Interaction

	SLOW SPEED	FAST SPEED	SPEED SIMPLE MAIN EFFECTS
Negative statements	6	6	0 (6 − 6 = 0)
Positive statements	8	8	0 (8 − 8 = 0)
Statement type simple main effects	2 (8 − 6 = 2)	2 (8 − 6 = 2)	
Averaging a treatment's simple main effects gives us the treatment's overall main effect:			
Simple main effect of ***Statement type*** in the slow presentation condition			2
Simple main effect of ***Statement type*** in the fast presentation condition			2
Average effect (overall main effect) of ***Statement type***			4/2 = 2
Simple main effect of SPEED in the negative statements condition			0
Simple main effect of SPEED in the positive statements condition			0
Average effect (overall main effect) of SPEED			0/2 = 0
Comparing a treatment's simple main effects tells us whether there is an interaction:			
Because there are no differences between statement type's two simple main effects (both are 2), there is no interaction. In other words, because the effect of statement type is not affected by the speed with which the statements are presented, there is no interaction.			

statements group scores an average of 2 points *higher* (8 − 6 = 2) than the negative statements group. Thus, there may be a simple main effect for statement type in the slow speed condition.

Looking at the second column shows you the effect of statement type for the fast-speed participants. In a way, looking at the second column is like looking at a simple experiment that manipulated statement type (while having all participants read the statements quickly). As you can see, the positive statement group scores an average of 2 points *higher* on the mood scale than the negative statement group (8 − 6 = 2). Thus, there may be a simple main effect for statement type in the fast-speed condition.

Because statement type increases mood for both the slow-speed and the fast-speed participants, there seems to be an overall main effect for statement type. Our best estimate of this average effect of statement type is that positive statements increase mood 2 points more than negative statements do.[4] Because statement type's effect does *not* differ depending on speed condition, there is *no* interaction between statement type and speed. Specifically, there is no interaction because positive statements increase mood by the same number

[4] Because of random error, you don't know what the effect actually is. Indeed, without using statistical tests, you can't claim that you have a significant main effect or an interaction. However, because our purpose in this section is to teach you how to interpret tables and graphs and because the tables and graphs you will see in journal articles will almost always be accompanied by a statistical analysis, we will pretend—in this section—that any differences between means are statistically significant and due entirely to treatment effects.

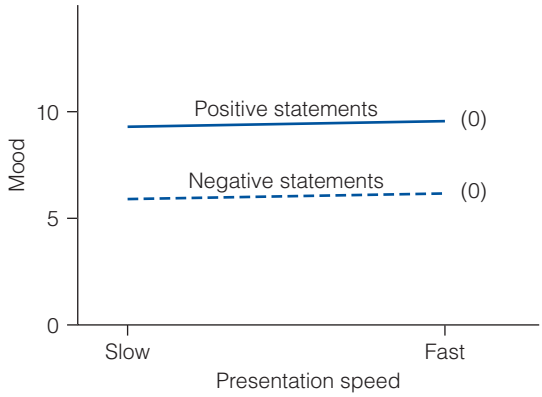

FIGURE **12.1** Main Effect for Statement Type, No Interaction

Note: Numbers in parentheses represent the speed simple main effects. Thus, the simple main effect of speed was 0 in both the positive statements condition and the negative statements condition.

of points (2) in the slow statements condition as they do in the fast statements condition.

Although making tables of means is a useful way to summarize data, perhaps the easiest way to interpret the results of a factorial experiment is to graph the means. To see how graphing can help you interpret your data, graph the data in Table 12.5. Before you plot your data, start by beginning to make a graph of a simple experiment that manipulates speed. Once you have a vertical *y*-axis labeled "Mood," and a horizontal *x*-axis that has labels for both slow presentation and fast presentation, you are ready to plot your data. Start by plotting two points representing the two means from the top row. Next, draw a line between those points and label that line "Negative statements." Then, plot the bottom row's two means. Draw a line between those two points and label that line "Positive statements." Your graph should look something like Figure 12.1. If it doesn't, please consult Box 12.1.

Figure 12.1 confirms what you saw in Table 12.5. Negative statements decreased mood relative to positive statements, as shown by the negative statements participants' line being below the positive statements participants' line. Speed did *not* affect mood, as shown by the fact that both lines stay perfectly level as they go from slow presentation (left) side to fast presentation (right) side of the graph.

Finally, there is no interaction between speed and statement type on mood, as shown by the fact that the lines are parallel.[5] The lines are parallel

[5] If you have a bar graph instead of a line graph, you can't simply look to see if the lines are parallel because there are no lines. Instead, the key is to see whether the relationship between the dark bar and the light bar on the left side of the graph is the same as the relationship between the dark bar and the light bar on the right side of the graph. For example, if, on the left side of the graph, the dark bar is taller than the light bar, but on the right side of the graph, the dark bar is shorter than the light bar, you may have an interaction. Alternatively, you may convert the bar graph into a line graph by (a) drawing one line from the top, right corner of the first dark bar to the top, left corner of the other dark bar, and (b) drawing a second line from the top, right corner of the first light bar to the top, left corner of the other light bar.

BOX 12.1 Turning a 2 × 2 Table Into a Graph

If you have never graphed a 2 × 2 before, you may need some help. How can you graph three variables (the two factors and the dependent variable) on a two-dimensional piece of paper? The short answer is that you need to use two lines instead of one.

To see how to make such a graph, get a sheet of notebook paper and a ruler. Starting near the left edge of the sheet, draw a 4-inch line straight down the page. This vertical line is called the *y*-axis. The *y*-axis corresponds to scores on the dependent measure. In this case, your dependent measure is mood. So, label the *y*-axis "Mood."

Now that you have a yardstick (the *y*-axis) for mood, your next step is to put marks on that yardstick. Having these marks will make it easier for you to plot the means accurately. Start marking the *y*-axis by putting a little hash mark on the very bottom of the *y*-axis. Label this mark "0." A half an inch above this mark, put another mark. Label the mark "5." Keep making marks until you get to "20."

Your next step is to draw a horizontal line that goes from the bottom of the *y*-axis to the right side of the page. (If you are using lined paper, you may be able to trace over one of the paper's lines.) The horizontal line is called the *x*-axis. On the *x*-axis, you should put one of your independent variables. It usually doesn't matter which independent variable you put on the *x*-axis. However, some people believe you should put the moderator variable on the x-axis. If you don't have a moderator variable, those same people believe you should put the factor you consider most important on the *x*-axis. For the sake of this example, put "Presentation speed" about an inch below the middle of the *x*-axis. Then, put a mark on the left-hand side of the *x*-axis and label this mark "Slow." Next, put a mark on the right side of the *x*-axis and label it "Fast."

You are now ready to plot the means in the first row of Table 12.5. Once you have plotted those two means, draw a straight line between those two means. Label that line "Negative statements." Next, plot the two means in the right column of Table 12.5. Then, draw a line between those two points. Label this second line (which should be above your first line) "Positive statements." Your graph should look something like Figure 12.1.

because speed is having the same effect on the positive statements group as it is on the negative statements group. In this case, speed is having no (0) effect on either group.

Note that if you graph your data, you need to see only whether the lines are parallel to know whether you have an interaction. *If your lines are parallel, you do not have an interaction.* If, on the other hand, your lines have different slopes, you may have an interaction.[6]

Instead of having no interaction and a main effect for statement type, you could have no interaction and a main effect for speed. This pattern of results is shown in Table 12.6. From the top row, you can see that in the negative statements groups, fast presentation increased mood by 5 points ($\underline{10} - 5 = 5$). Looking at the bottom row, you see that in the positive statements groups, fast presentation also increased mood scores by 5 points ($\underline{\mathbf{10}} - \mathbf{5} = 5$). By averaging the effect of speed over both the negative statements and the positive statements conditions, you could estimate that speed's average effect, the overall main effect of speed, was 5.

Whereas looking at the rows tells you about the effects of speed, looking at the columns tells you about the effect of statement type. Looking at the

[6] Remember that because of random error, we don't know what the effect actually is. To know whether we had an interaction, we would need to do a statistical significance test.

TABLE **12.6**
Main Effect for Speed, No Interaction

	SLOW SPEED	FAST SPEED	SPEED SIMPLE MAIN EFFECTS
Negative statements	5	10	5 (10 − 5 = 5)
Positive statements	5	10	5 (10 − 5 = 5)
Statement type simple main effects	0 (5 − 5 = 0)	0 (10 − 10 = 0)	

Averaging a treatment's simple main effects gives us the treatment's overall main effect:

Simple main effect of ***Statement type*** in the slow presentation condition	0
Simple main effect of ***Statement type*** in the fast presentation condition	0
Average effect (overall main effect) of ***Statement type***	0/2 = 2
Simple main effect of SPEED in the negative statements condition	5
Simple main effect of SPEED in the positive statements condition	5
Average effect (overall main effect) of SPEED	10/2 = 5

Comparing a treatment's simple main effects tells us whether there is an interaction:

Because there are no differences between statement type's two simple main effects (both are 0), there is no interaction. In other words, because the effect of statement type is not affected by the speed with which the statements are presented, there is no interaction.

first column tells you about the effect of statement type in the slow presentation conditions. In the slow presentation conditions, the negative statement participants were in the same mood as the positive statements participants (both averaged 5 on the mood scale). Thus, there was no simple main effect of statement type in the slow presentation conditions.

Looking at the second column (the fast presentation column) tells you about the effect of statement type in the fast conditions. You can see that, in the fast presentation condition, the negative statement participants were in the same mood as positive statements participants (both averaged 10 on the mood scale). Thus, there was no simple main effect for statement type in the fast presentation condition.

To determine the overall main effect of statement type, compute the average of the two statement type simple main effects. Because there was no (zero) observed effect for varying statement type in both the slow presentation condition (the first column) and the fast presentation condition (the second column), there is no (zero) overall main effect for varying statement type.

To determine whether there is a statement type × speed interaction, you could subtract the statement type simple main effects from each other (0 − 0 = 0). Or, you could subtract the speed simple main effects from each other (5 − 5 = 0). Either way, the result is zero, suggesting that you don't have a speed × statement type interaction. You do not have an interaction because the effect of speed is not affected by the statement type variable: Increasing presentation speed increases mood by 5 points, regardless of whether statements are positive or negative.

TABLE **12.7**

Main Effect for Speed and Statement Type, No Interaction

	SLOW SPEED	FAST SPEED	SPEED SIMPLE MAIN EFFECTS
Negative statements	4	8	4 (8 − 4 = 4)
Positive statements	6	10	4 (10 − 6 = 4)
Statement type simple main effects	2 (6 − 4 = 2)	2 (10 − 8 = 2)	

Averaging a treatment's simple main effects gives us the treatment's overall main effect:

Simple main effect of ***Statement type*** in the slow presentation condition	2
Simple main effect of ***Statement type*** in the fast presentation condition	2
Average effect (overall main effect) of ***Statement type***	4/2 = 2
Simple main effect of SPEED in the negative statements condition	4
Simple main effect of SPEED in the positive statements condition	4
Average effect (overall main effect) of SPEED	8/2 = 4

Comparing a treatment's simple main effects tells us whether there is an interaction:

Because there are no differences between statement type's two simple main effects (both are 2), there is no interaction. In other words, because the effect of statement type is not affected by the speed with which the statements are presented, there is no interaction.

Two Main Effects and No Interaction

Table 12.7 reflects another pattern of effects you might obtain. From the first row, you can see that, in the negative statements groups, fast statements increased mood scores by 4 points (8 − 4). Looking at the second row, you see that, in the positive statements groups, speed also increased mood scores by 4 points (10 − 6). Averaging the effect of speed over all the statement type conditions, you find that the average effect of speed (the overall main of speed) was to increase mood scores by 4 points.

Looking at the columns tells you about the effect of varying statement type. The first column tells you about what happens in the slow presentation conditions. As you can see, in the slow presentation conditions, the participants who read positive statements averaged 2 points higher (6 − 4) on the mood scale than those who read negative statements. Looking at the second column, you see that, in the fast presentation conditions, participants who read positive statements score, on the average, 2 (10 − 8) points higher on the mood scale than participants who read negative statements. Because positive statements increase mood in both the slow presentation and the fast presentation groups, it appears that there is a statement type main effect.

Comparing the two columns tells you that there is *no* interaction because the effect of statement type is unaffected by speed. As Table 12.7 demonstrates, the effect of statement type is independent of (does not depend on) speed. In this case, positive statements increase mood by 2 points, regardless of whether participants are in the slow or fast thought condition.

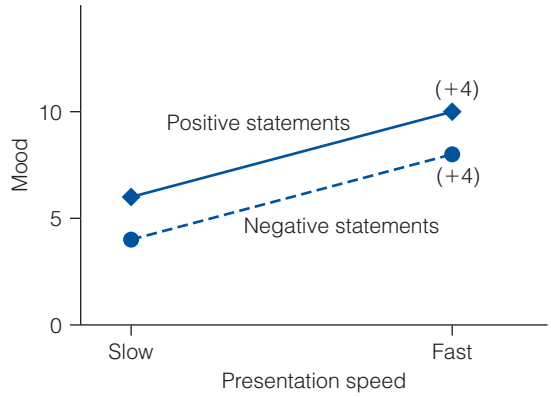

FIGURE **12.2** Main Effect for Statement Type and Speed, No Interaction

Note: Numbers in parentheses represent the speed simple main effects. Thus, the simple main effect of speed was +4 in both the positive statements condition and in the negative statements condition.

To look at this lack of statement type × speed interaction from a different perspective, look at the rows. Comparing the rows shows you that the effect of speed is unaffected by the type (positive or negative) of statement. Specifically, fast statements increase mood by 4 points for both the negative statement groups and the positive statements groups.

We have shown you two ways to use a table of means (like Table 12.7) to determine whether you have an interaction: (1) by comparing (subtracting) the simple main effects of the two rows, and (2) by comparing (subtracting) the simple main effects of the two columns. There is a third way. If either simple main effect for a factor is the same as that factor's overall main effect, you do *not* have an interaction. Thus, in the current example, we know there is no interaction because the simple main effect of fast statements in the positive statements conditions (4) is the same as the overall main effect of fast statements (4).

Although a table of means gives you valuable information, you may understand your data better if you graph the means. To appreciate this point, look at a graph of Table 12.7's means: Figure 12.2. As you can see from the negative statements line being below the positive statements line, positive statements increased mood relative to negative statements. As you can see from both lines sloping upward as they go from the slow statements (left) side to fast statements (right) side of Figure 12.2, fast statements, relative to slow statements, increased mood. Finally, as you can see from the parallel lines, there is no interaction between speed and statement type. The lines are parallel because speed affects the negative statements groups the same (parallel) way that it affects the positive statements groups.

Two Main Effects and an Interaction

Now imagine that you got a very different set of results from your statement type–speed study. For example, suppose you found the results in Table 12.8.

As the table shows, you have main effects for both speed and statement type. The average effect of fast statements is to *decrease* mood scores by

TABLE **12.8**

Main Effect for Speed and Statement Type, and a (Crossover) Interaction

	SLOW SPEED	FAST SPEED	SPEED SIMPLE MAIN EFFECTS
Negative statements	10	12	2 (12 − 10 = 2)
Positive statements	**20**	**12**	**−8 (12 − 20 = −8)**
Statement type simple main effects	10 (**20** − 10 = 10)	0 (**12** − 12 = **0**)	

Averaging a treatment's simple main effects gives us the treatment's overall main effect:

Simple main effect of ***Statement type*** in the slow presentation condition	10
Simple main effect of ***Statement type*** in the fast presentation condition	0
Average effect (overall main effect) of ***Statement type***	10/2 = 5
Simple main effect of SPEED **in the negative statements condition**	2
Simple main effect of SPEED **in the positive statements condition**	**−8**
Average effect (overall main effect) of SPEED	−6/2 = −3

Comparing a treatment's simple main effects tells us whether there is an interaction:

Because there are differences between statement type's two simple main effects (one is 10, one is 0), there is an interaction. In other words, because the effect of statement type is affected by the speed with which the statements are presented, there is an interaction.

3 points, and the average effect of positive statements is to *increase* mood scores by 5.

Although, on the average, fast statements have an effect, the specific effect of fast statements varies depending on whether participants read negative or positive statements. In the positive statements condition, fast statements, relative to slow statements, *increased mood* by 2 points (12 vs. 10). In the negative statements condition, on the other hand, fast statements *decreased* mood by 8 points (**12 vs. 20**). Because the effect of speed differs depending on statement type, there is an interaction.

To see this interaction, look at Figure 12.3a. As you can see, the lines are not parallel because the slope of the negative statements line is different from the slope of the positive statements line. This difference in slope indicates that the effect of speed is different for the negative statements groups than for the positive statements groups. In this case, the negative statements line slopes upward (indicating that negative statements participants are in a *better* mood in the fast statements condition than in the slow statements condition), whereas the positive statements line slopes downward (indicating that positive statements participants are in a *worse* mood in the fast condition than in the slow condition). When the lines slope in opposite directions—indicating that the effect a treatment has with one group of participants is opposite from that treatment's effect on the other group of participants—the interaction is often called a **crossover (disordinal) interaction** (because the lines often *cross*).

Crossover interactions are also called *disordinal interactions* because they can't be merely the result of having ordinal rather than interval data. That is,

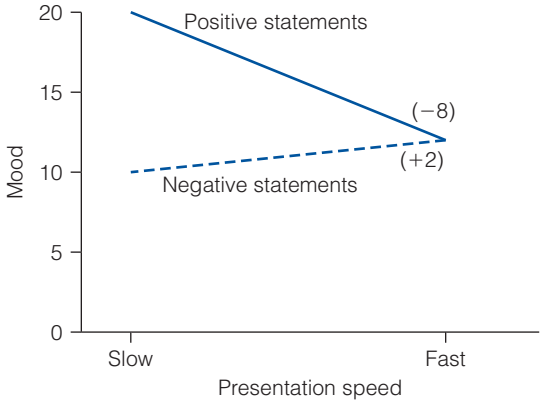

FIGURE 12.3a Main Effects for Statement type and Speed, and a Crossover (Disordinal) Interaction

Note: Numbers in parentheses represent the speed simple main effects. Thus, the simple main effect of speed was −8 in the positive statements condition but +2 in the negative statements condition.

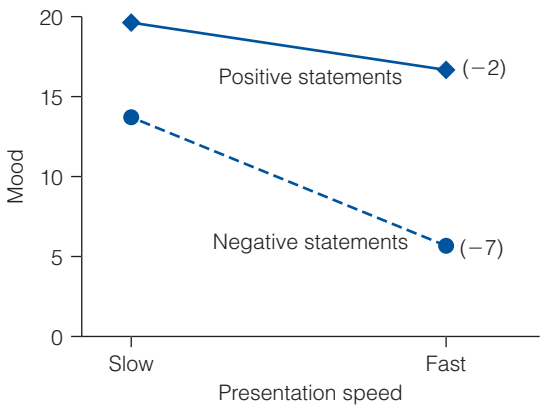

FIGURE 12.3b Main Effects for Statement type and Speed and an Ordinal Interaction

Note: Numbers in parentheses represent the speed simple main effects. Thus, the simple main effect of speed was only −2 in the positive statements condition but was −7 in the negative statements condition.

even if your measure can't tell you how much more of a quality one participant has than another, that problem with your measure won't make it look like the treatment is increasing the quality in one condition but decreasing it in the other condition.

Such a measurement problem, however, could cause other types of interactions. To see how, consider Figure 12.3b, in which both lines slope downward but the negative statements line slopes downward more sharply than the positive statements line. As you can see from Figure 12.3b, the lines are not parallel—and, therefore, there is an interaction. Such an interaction could be due to the negative statements participants being more affected by the fast thought manipulation than the positive statements participants were. Although such an interaction could be due to the treatment having more of an effect in one condition than in another, such an interaction could also be due to an ordinal measure creating the *illusion* that the treatment has more of an effect in one condition than the other. For example, suppose the mood score was based on participants selecting the adjective that best describes them. If checking "omnipotent" is scored as "20," checking "superior" is scored as "18," checking "powerful" is scored as "12," and checking "influential" is scored as "7," this measure may be ordinal. With such an ordinal measure, although going from 20 to 18 is clearly less of a decrease in *measured* mood than going from 12 to 7, going from 20 to 18 (from omnipotent to merely superior) may *not* be less of a difference in *actual* mood than going from 12 to 7 (from powerful to influential). Because interactions that *appear* to be due to a treatment having *more* of an effect in one condition than in another could actually be an illusion caused by having ordinal data, such interactions are called *ordinal interactions*.

TABLE **12.9**

No Main Effects for Speed or Statement With a (Crossover) Interaction

	SLOW SPEED	FAST SPEED	SPEED SIMPLE MAIN EFFECTS
Negative statements	10	15	+5 (15 − 10 = 5)
Positive statements	15	10	−5 (10 − 15 = −5)
Statement type simple main effects	+5 (15 − 10 = 5)	−5 (10 − 15 = −5)	

Averaging a treatment's simple main effects gives us the treatment's overall main effect:

Simple main effect of ***Statement type*** in the slow presentation condition	+5
Simple main effect of ***Statement type*** in the fast presentation condition	−5
Average effect (overall main effect) of ***Statement type***	0/2 = 0
Simple main effect of SPEED in the negative statements condition	+5
Simple main effect of SPEED in the positive statements condition	−5
Average effect (overall main effect) of SPEED	0/2 = 0

Comparing a treatment's simple main effects tells us whether there is an interaction:

Because there are differences between statement type's two simple main effects (one is +5, one is −5), there is an interaction. In other words, because the effect of statement type *depends on* the speed with which the statements are presented, there is an interaction.

An Interaction and No Main Effects

You have seen that you can have main effects with interactions, but can you have interactions without main effects? To answer this question, consider the data in Table 12.9 and Figure 12.4a.

From the graph (Figure 12.4a), you can see that the lines are not parallel. Instead, the lines actually cross. In this case, the crossover interaction is due to speed having one kind of effect (increasing mood) in the negative statements condition, but having an opposite effect (decreasing mood) in the positive statements condition. (In this case, "X" marks the crossover interaction. However, graphs of crossover interactions don't always look like *X*s. As you can see from Figure 12.4b, a graph of a crossover interaction sometimes looks like a sideways "V" rather than an "X.")

Although you have an interaction between statement type and speed, you do not have a main effect for either statement type or speed. As you can tell by looking at Table 12.9, the slow presentation groups have the same average mood as the fast presentation groups. Therefore, there isn't a speed main effect. Similarly, because the negative statements groups have the same average mood as the positive statements groups, there isn't a statement type main effect.

Thus, you would have to say that neither statement type nor speed has a main effect. Yet, you would not want to say that neither statement type nor speed has any effect. Instead, you would either say that (a) statement type has an effect, but its effect *depends* on the speed at which the statements are

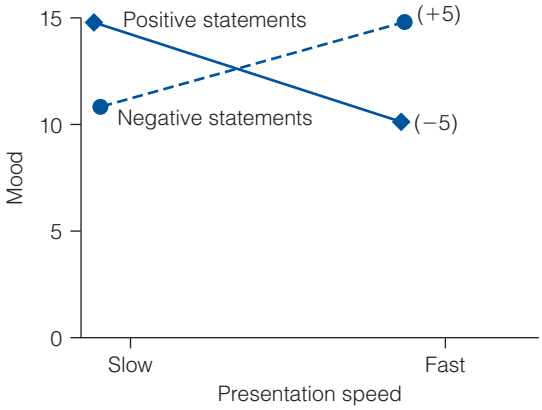

FIGURE **12.4a** No Main Effects and a Crossover Interaction: The Classic "X"-Shaped Pattern

Note: Numbers in parentheses represent the speed simple main effects. Thus, the simple main effect of speed was −5 in the positive statements condition but +5 in the negative statements condition.

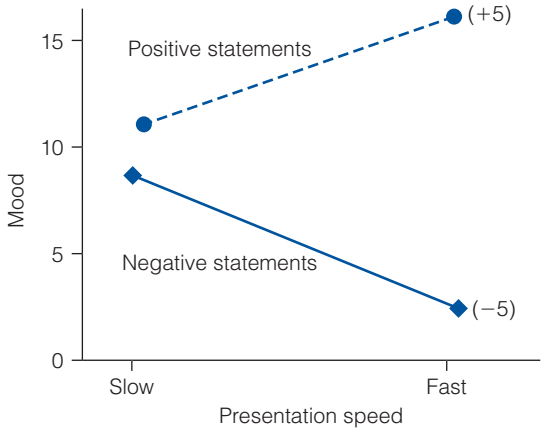

FIGURE **12.4b** No Main Effects and a Crossover Interaction: The Classic "V"-Shaped Pattern

Note: Numbers in parentheses represent the speed simple main effects. Thus, the simple main effect of speed was +5 in the positive statements condition but −5 in the negative statements condition.

presented, or (b) speed has an effect, but its effect *depends* on whether the statements are positive or negative.

Regardless of whether you emphasize the effect of statement type (as in the first statement) or the effect of speed (as in the second statement), you cannot talk about the effect of one variable without talking about the other. In short, if you have an interaction, the effect of one variable depends on the other—even when you don't have any main effects.

An Interaction and One Main Effect

You have seen that you can have no main effects and an interaction. You have also seen that you can have two main effects and an interaction. Can you also have one main effect and an interaction? Yes—such a pattern of results is listed in Table 12.10 and graphed in Figure 12.5.

As Table 12.10 reveals, the average effect of varying statement type is zero. (The –2 effect of statement type in the slow condition is cancelled out by the +2 effect of statement type in the fast condition.) The average effect of varying speed, on the other hand, is to increase mood scores by 2. Note, however, that speed's effect is uneven. In the negative statements condition, fast statements have no observable effect ($10 - \underline{10} = 0$). But in the positive statements condition, speed has an effect ($\underline{12} - 8 = 4$). Because the effect of speed differs depending on statement type, there is a speed × statement type interaction.

Figure 12.5 tells the same story. By looking at that figure, you realize there may be an interaction because the lines are not parallel. They are not parallel because the effect of speed is dramatic in the positive statements conditions but undetectable in the negative statements conditions.

TABLE 12.10
Main Effect for Speed With an Interaction

	SLOW SPEED	FAST SPEED	SPEED SIMPLE MAIN EFFECTS
Negative statements	10	10	0 (10 − 10 = 0)
Positive statements	8	12	4 (12 − 8 = 4)
Statement type simple main effects	−2 (8 − 10 = −2)	+2 (12 − 10 = +2)	

Averaging a treatment's simple main effects gives us the treatment's overall main effect:

Simple main effect of ***Statement type*** in the slow presentation condition	−2
Simple main effect of ***Statement type*** in the fast presentation condition	+2
Average effect (overall main effect) of ***Statement type***	0/2 = 0
Simple main effect of Speed in the negative statements condition	0
Simple main effect of Speed in the positive statements condition	4
Average effect (overall main effect) of Speed	4/2 = 2

Comparing a treatment's simple main effects tells us whether there is an interaction:

Because there are differences between statement type's two simple main effects (one is −2, the other is +2), there is an interaction. In other words, because the effect of statement type is affected by the speed with which the statements are presented, there is an interaction.

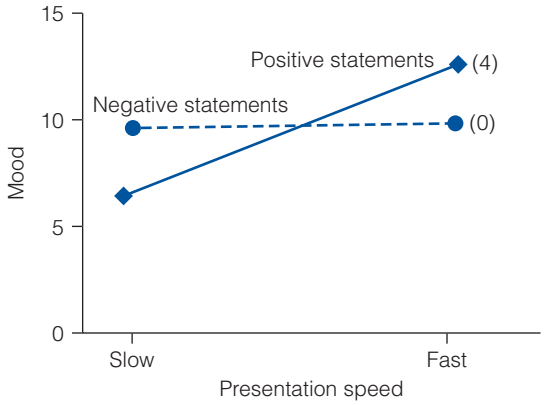

FIGURE 12.5 Main Effect for Speed With an Interaction

Note: Numbers in parentheses represent the speed simple main effects. Thus, the simple main effect of speed was 4 in the positive statements condition but 0 in the negative statements condition. Because the simple main effect of speed differs depending on statement type, there is an interaction.

TABLE **12.11**
No Main Effects and No Interaction

	SLOW SPEED	FAST SPEED
Negative statements	12	12
Positive statements	12	12

Whereas you can glance at Figure 12.5 and instantly see the interaction, seeing the main effects requires more mental visualization. If there is a main effect for statement type, one of the statement type lines should, on the average, be higher than the other. When one line is always above the other, it is easy to tell whether there seems to be a main effect. In this case, however, the lines cross—making it hard to tell whether one line is, on the average, above the other. If you get a ruler and mark the midpoint of each line, you will see that the midpoint of both lines is at the same spot. Or, you may realize that the negative statements line is below the positive statements line just as often and to the same extent as it is above the positive statements line. In either case, you would conclude that there is no main effect for statement type.

To determine whether there is a main effect for speed, you could mentally combine the two lines. If you do that, you would "see" that this combined line slopes upward, indicating a positive main effect for speed. (If you can't visualize such a line, you can create one in three steps. First, take a ruler and put a point halfway between the left ends of the two lines [i.e., a point halfway between the two slow statements points]. Second, put a point halfway between the right ends of the two lines [i.e., a point halfway between the two fast statements points]. Third, draw a line between the two points you just drew.) Alternatively, you could reason that because the positive statements line slopes upward and the negative statements line stays level, the average of the two lines would be to slope upward.

If you prefer not to think about lines at all, convert the graph into a table of means. To practice, take Figure 12.5 and see if you can convert it into a table resembling Table 12.10. Once you have your table of means, you will be able to see that the average for the fast statements groups is higher than the average for the slow statements groups.

No Main Effects and No Interaction

The last pattern of results you could obtain is to get no statistically significant results. That is, you could fail to find a statement type effect, fail to find a speed effect, and fail to obtain an interaction between statement type and speed. An example of such a dull set of findings (possibly caused by a lack of power) is listed in Table 12.11.

ANALYZING RESULTS FROM A FACTORIAL EXPERIMENT

You can now graph and describe all the possible patterns of results from a 2×2 experiment. But how would you analyze your results to determine whether a main effect or an interaction is significant?

You would probably use analysis of variance (ANOVA) to analyze your data. Using ANOVA to analyze a factorial experiment is similar to using ANOVA to analyze data from a single factor experiment. The main difference is that instead of testing for one main effect, you will be testing for two main effects and an interaction. Thus, your ANOVA summary table might look like this:

SOURCE OF VARIANCE	SUM OF SQUARES (*SS*)	*df*	MEAN SQUARE (*MS*)	*F*
Speed Main Effect (*A*)	900	1	900	9.00
Statement type Main Effect (*B*)	200	1	200	2.00
Interaction ($A \times B$)	100	1	100	1.00
Error Term (within groups)	3600	36	100	
Total	4800	39		

What Degrees of Freedom Tell You

Despite the fact that this ANOVA table has two more sources of variance than the ANOVA for the multiple-group experiment described in Chapter 11, most of the rules that apply to the ANOVA table for that design also apply to the table for a factorial design (see Box 12.2). In terms of degrees of freedom, you can still use the two rules we discussed in Chapter 11:

1. The number of treatment levels is one more than the treatment's degrees of freedom. Because the ANOVA summary table above states that the degrees of freedom for speed is 1, we know that the study used two levels of speed. Likewise, because the degrees of freedom for statement type is 1, we know the study used two statement types. Thus, the ANOVA summary table tells us that the study used a 2×2 design.
2. The total number of participants is one more than the total degrees of freedom. Therefore, because the ANOVA table states that the total degrees of freedom was 39, we know that there were 40 (39 + 1) participants in the experiment.

The only new rule is for the interaction's degrees of freedom. To calculate the interaction term's degrees of freedom, multiply the degrees of freedom for the main effects making up that interaction. For a 2×2 experiment, that would be 1 (*df* for first main effect) $\times$ 1 (*df* for second main effect) = 1. For a 2×3 experiment, that would be 1 (the *df* for the first main effect would be 1) $\times$ 2 (the *df* for the second main effect) = 2.

What *F* and *p* Values Tell You

To determine whether an effect was significant, you look at the *p* value for the effect. If the *p* value is less than .05, the effect is statistically significant. If you do not have the *p* values, compare the *F* for that effect to the value given in the *F* table (see Table 3 in Appendix F) under the appropriate number of degrees of freedom. If your obtained *F* is larger than the value in the table, the effect is statistically significant.

BOX 12.2 The Mathematics of an ANOVA Summary Table for Between-Subjects Factorial Designs

1. Degrees of freedom (*df*) for a main effect equal 1 less than the number of levels of that factor. If there are 3 levels of a factor (low, medium, high), that factor has 2 *df*.
2. Degrees of freedom for an interaction equal the product of the *df* of the factors making up that effect. If you have an interaction between a factor that has 1 *df* and a factor that has 2 *df,* that interaction has 2 *df* (because $1 \times 2 = 2$).
3. To get the total degrees of freedom, subtract 1 from the number of participants. Therefore, if you have 60 participants, the total degrees of freedom should be 59 (60–1).
4. To get the *df* for the error term, determine how many groups you had. Then, subtract the number of groups from the number of participants. In a 2×2, you have 4 (2×2) groups. Therefore, if you had 60 participants, your *df* error is 56 (60–4). If you had a 3×2, you would have 6 (3×2) groups. Therefore, the *df* error would be 54 (60–6). Another way to get the *df* error is to (a) add up the *df* for all the main effects and interactions, and then (b) subtract that sum from the total degrees of freedom. Thus, if you had 1 *df* for the first main effect, 1 *df* for the second main effect, 1 *df* for the interaction, the sum of the *df* for your main effects and interactions would be **3** ($1 + 1 + 1$). You would then subtract that sum (3) from the *df* total. Thus, if the *df* total was 59, your error term would be 56 (59–3).
5. To get the mean square for any effect, get the sum of squares for that effect, and then divide by that effect's *df*. If an effect's sum of squares was 300, and its *df* was 3, its mean square would be 100 (because $300/3 = 100$). If the effect's sum of squares was 300, and its *df* was 1, its mean square would be 300 (because $300/1 = 300$).
6. To get the *F* for any effect, get its mean square and divide it by the mean square error. If an effect's mean square was 100, and the mean square error was 50, the *F* for that effect would be 2 (because $100/50 = 2$).

What Main Effects Tell You: On the Average, the Factor Had an Effect

Usually, you will want to start your inspection of the *ANOVA* results by seeing whether any of your overall main effects are significant. If you have a significant effect for a factor, the overall effect of that factor is either to increase or to decrease scores on the dependent measure. If you have a significant main effect, your next step would be to find out whether this main effect is qualified by an interaction.

If the interaction was not significant, your conclusions are simple and straightforward. Having no interactions means there are no "ifs" or "buts" about your main effects. That is, you have not found anything that would lead you to say that the main effect occurs only under certain conditions. For instance, if you have a main effect for statement type and no interactions, statement type had the same kind of effect throughout your experiment—no matter the speed at which participants read those statements. When you don't have interactions, you can just talk about the overall main effects. Thus, your Results section might resemble the following:

> A 2 (Statement type: positive statements, negative statements) × 2 (Speed: slow, fast) between-subjects *ANOVA* was conducted to assess the effects of statement type and speed on mood. Contrary to our hypothesis, this analysis did not find

that the positive statements group was in a better mood ($M = 11.8$) than the negative statements group ($M = 12.2$), $F(1, 48) = 2.14$, *ns*. However, the analysis did reveal the expected main effect for speed, with participants in the fast thought groups scoring higher on mood ($M = 16$) than participants in the slow thought groups ($M = 8$), $F(1, 48) = 4.21$, $p = .04$, $r_{\text{effect size}} = .12$. The speed main effect was not qualified by a speed × statement type interaction, $F(1, 48) = 1.42$, *ns*.

If, on the other hand, you had an interaction, you would replace the last sentence with something like the following:

These findings are qualified, however, by a significant speed × statement type interaction, $F(1, 48) = 4.60$, $p = .04$, $\eta^2 = .08$. In the positive statements conditions, the participants in the slow presentation condition scored almost as high on the mood scale ($M = 16.1$, $SD = 3.33$) as participants in the fast presentation condition ($M = 16.3$, $SD = 3.46$). However, in the negative statements conditions, participants in the slow presentation condition were in a worse mood ($M = 6.11$, $SD = 3.11$) than participants in the fast presentation condition ($M = 10.1$, $SD = 3.22$).

What Interactions Usually Tell You: Combining Factors Leads to Effects That Differ From the Sum of the Individual Main Effects

As you just saw, when you have a significant interaction, describing the results is more complicated than when you don't have a significant interaction. You can't just talk about one variable's effect without also stating that the variable's effect depends on (is moderated by, is qualified by) a second variable.

At a more concrete level, having an interaction means that a treatment factor has a different effect on one group of participants than on another. In our statement type–speed example, having an interaction would mean that the simple main effect of statement type in the slow statements condition is different from the simple main effect of statement type in the fast statements condition. In that case, because statement type's simple main effects would differ, rather than talking only about statement type's general, average, overall main effect, you would talk about the specific, individual, simple main effects that make up that overall main effect.

Before you can talk about those simple main effects, however, you must understand them. The easiest way to understand the pattern of the simple main effects—and thus understand the interaction—is to graph them.[7] In addition to looking at the slope of each line, examine the relationship between your lines to see why they aren't parallel.

If the lines are sloping in different directions, you have a disordinal interaction and you know that the interaction is not merely an artifact of having ordinal data. Therefore, you know that the treatment has one effect in one condition and a different effect in another.

If, on the other hand, both lines are sloping in the same direction but one is steeper than the other, you have an ordinal interaction and you know that

[7] Interactions suggest that, rather than looking at the overall main effects, you should look at the individual simple main effects. One way to understand an interaction is to do statistical analyses on the individual simple main effects. The computations for these tests are simple. However, there are some relatively subtle issues involved in deciding which test to use.

your interaction may merely be an artifact of having ordinal data. Therefore, you can't be confident that the interaction is due to the treatment having a stronger effect on one group than on another.

PUTTING THE 2 × 2 FACTORIAL EXPERIMENT TO WORK

You now understand the logic behind the 2 × 2 design. In the next sections, you will see how you can use the 2 × 2 to produce research that is more interesting, has greater construct validity, and has greater external validity than research produced by a simple experiment.

Looking at the Combined Effects of Variables That Are Combined in Real Life

Suppose you are aware of research showing that driving while talking on cell phones impairs driving performance and that you are aware that driving while drunk impairs driving performance, but you are unaware of any research looking at the combined effects of both these factors. Then, if you think a study examining both factors would have practical implications (some people use cell phones while driving drunk) or theoretical implications (to see whether inattention is the mechanism for both), you might propose a study that looked at both factors at once (you would use a driving simulator rather than having people actually drive). Similarly, you could look at how driving performance was affected by the interaction of cell phone use with any of the following variables: sleep deprivation, caffeine, number of passengers in the car, or driving conditions.

Ruling out Demand Characteristics

Suppose you design a simple experiment in which half of your participants think about their own death and the other half think about going to the dentist. You expect that participants made to think about death are more likely to have happy thoughts than people made to think about going to the dentist. A friend criticizes your proposal, suggesting that your findings would just be the result of participants playing along with your hypothesis. To test that possibility, you could add two more groups to your study: a group that imagines how they would feel if they were in the death-salience condition and a group that imagines how they would feel if they were in the dental-pain condition (you are now proposing a replication of DeWall & Baumeister, 2007). If the pattern of results for the groups that really experienced the treatment is different from the pattern of results for the groups that role-played receiving the treatment, you would show that your hypothesis was not as intuitive as your friend believed. Note that all simple experiments involve comparing two levels of treatment (e.g., treatment 1 vs. treatment 2), and that you could convert most of those experiments into 2 (treatment 1 vs. treatment 2) × 2 (imagined vs. direct experience) experiments just by adding two groups that imagine—rather that actually—experience the treatments.

Adding a Replication Factor to Increase Generalizability

The generalizability of results from a single simple experiment can always be questioned. Critics ask questions such as, "Would the results have been

different if a different experimenter had performed the study?" and "Would the results have been different if a different manipulation had been used?" Often, the researcher's answer to these critics is to do a **systematic replication**: a study that varies from the original only in some minor aspect, such as using different experimenters or different stimulus materials.

For example, Morris (1986) found that students learned more from a lecture presented in a rock-video format than from a conventional lecture. However, Morris used only one lecture and one rock video. Obviously, we would have more confidence in his results if he had used more than one conventional lecture and one rock-video lecture.

Morris would have benefited from doing a 2 × 2 experiment. Because the 2 × 2 factorial design is like doing two simple experiments at once, Morris could have (1) obtained his original findings and (2) replicated them with a different set of stimulus materials. Specifically, in addition to manipulating the factor of presentation type (conventional lecture vs. rock-video lecture), he could also have manipulated the replication factor of **stimulus sets**: the particular stimulus materials shown to one or more groups of participants. For example, he could have done a 2 (presentation type [conventional lecture vs. rock-video format]) × 2 (stimulus sets [material about Shakespeare vs. material about economics]) study. Because psychologists often want to show that the manipulation's effect can occur with more than just one particular stimulus set, experimenters routinely include stimulus sets as a replication factor in their experiments.[8]

Stimulus sets are not the only replication factor that researchers use. Some researchers employ more than one experimenter to run the study and then use experimenter as a factor in the design.

Some of these researchers use experimenter as a factor to show the generality of their results. Specifically, they want to show that certain experimenter characteristics (gender, attractiveness, status) do not alter the treatment's effect.

Other researchers use experimenters as a factor to establish that the experimenters are not biasing the results. For instance, Ranieri and Zeiss (1984) were worried that experimenters might unintentionally influence participants' responses to their experiment's dependent measure: a self-report scale of mood. Therefore, they used three experimenters and randomly assigned participants to experimenter. If different experimenters had obtained different patterns of results, Ranieri and Zeiss would have suspected that the results might be due to experimenter effects rather than to the manipulation itself.

Thus far, we have discussed instances in which the investigator's goal in using the factorial design was to increase the generalizability of the experimental results. Thus, in a study that uses stimulus set as a replication factor, researchers hope that the treatment × stimulus set interaction will not be significant. Similarly, most researchers who use experimenter as a factor hope that there will not be a treatment × experimenter interaction.

---

[8] However, psychologists have not all agreed that the traditional, fixed-effects analysis of variance should be used to analyze such studies (see Clark, 1973; Cohen, 1976; Coleman, 1979; Kenny & Smith, 1980; Richter & Seay, 1987; Wickens & Keppel, 1983; Wike & Church, 1976).

Using an Interaction to Find an Exception to the Rule: Looking at a Potential Moderating Factor

Often, however, researchers are interested in finding an interaction. For example, you may read about a study's results and say to yourself, "But I bet that would not happen under _____ conditions." In that case, you should do a study in which you essentially repeat the original experiment except that you add what you believe will be a moderating factor that will interact with the treatment.

To see how a moderating factor experiment would work, let's look at a study by Jackson and Williams (1985). Although aware of the phenomenon of social loafing—individuals don't work as hard on tasks when they work in groups as when they work alone—Jackson and Williams felt that social loafing would not occur on extremely difficult tasks. Therefore, they did a study, which, like most social-loafing studies, manipulated whether participants worked alone or in groups. In addition, they added what they thought would be a moderating factor—whether the task was easy or difficult (e.g., whether participants completed a simple maze or a challenging maze).

As expected, and as other studies had shown, social loafing occurred. But, social loafing occurred only when the task was easy. When the task was difficult, the reverse of social loafing occurred: Participants worked harder in groups than alone. This interaction between task difficulty and number of workers confirmed Jackson and Williams's hypothesis that task difficulty moderated social loafing (see Figure 12.6).

To see how you could take advantage of Jackson and Williams's research strategy, let's review what they did. With part of their study, they replicated an existing finding (the social-loafing main effect). With the other part, they tested whether another variable would moderate (interact with) the social-loafing main effect. If you like this strategy of proposing a study that tests both a safe prediction (e.g., a replication) and a risky prediction (e.g., an untested interaction), consider a moderating factor study. Note that this strategy works well if you have an idea about how to neutralize a bad effect

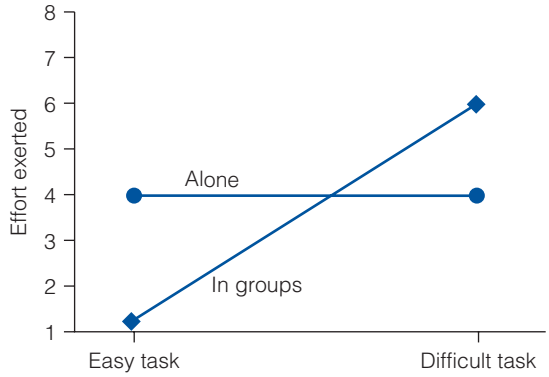

FIGURE **12.6** Interaction Between Task Difficulty and Number of Coworkers on Effort

Note: Effort was scored on a 1-to-7 scale, with higher numbers indicating more effort.

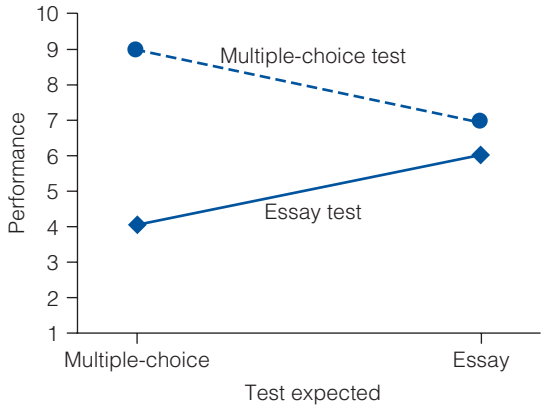

FIGURE **12.7** The Effect of Expectations and Type of Test on Performance

(e.g., a training program that would reduce frustration's effect of increasing aggression) or intensify a good effect (e.g., instructions that may improve the positive effects of a placebo). For more tips on designing a moderating factor study, see Chapter 3.

Using Interactions to Create New Rules

Although we have discussed looking for an interaction to find an exception to an existing rule, some interactions do more than complicate existing rules. Some interactions reveal new rules. Consider Tversky's (1973) 2 × 2 factorial experiment. She randomly assigned students to one of four conditions:

1. Student expected a multiple-choice test and received a multiple-choice test.
2. Student expected a multiple-choice test and received an essay test.
3. Student expected an essay test and received a multiple-choice test.
4. Student expected an essay test and received an essay test.

She found an interaction between type of test expected and test received. Her interaction showed that participants did better when they got the *same* kind of test they expected (see Figure 12.7).

Similarly, a researcher might find an interaction between mood (happy, sad) at the time of learning and mood (happy, sad) at the time of recall. The interaction might reveal that recall was best when participants were in the *same* mood at the time of learning as they were at the time of recall. As you can see, the 2 × 2 experiment may be useful for you if you are interested in assessing the effects of *similarity*.

Conclusions About Putting the 2 × 2 Factorial Experiment to Work

As you have seen, expanding a simple experiment into a 2 × 2 experiment allows you to test more—and more interesting—hypotheses. You can look at the main effect of the factor you would have studied with the simple experiment, plus the main effect of an additional factor, plus the interaction

between those two factors. In many cases, the hypothesis involving the interaction may be the most interesting.

HYBRID DESIGNS: FACTORIAL DESIGNS THAT ALLOW YOU TO STUDY NONEXPERIMENTAL VARIABLES

Rather than converting a simple experiment into a 2 × 2 experiment by adding a second experimental factor, you could convert a simple experiment into a 2 × 2 hybrid design by adding a nonexperimental factor. The nonexperimental factor could be any variable that you cannot randomly assign, such as age, gender, or personality type.

Hybrid Designs' Key Limitation: They Do Not Allow Cause–Effect Statements Regarding the Nonexperimental Factor

In such a hybrid 2 × 2 design, you could make cause–effect statements about the effects of the experimental factor, but *you could not make any cause–effect statements regarding the nonexperimental factor*. Thus, although the study described in Table 12.12 includes gender of participant as a variable, the study does not allow us to say anything about the effects of a participant's gender.

You can't make cause–effect (causal) statements regarding the effects of the participant's gender because your two groups may differ not only in terms of gender but also in hundreds of other ways. For example, they may differ in terms of college major, age, self-esteem, religiosity, parental support, or loneliness. Any one of the hundreds of potential differences between the groups might be responsible for the difference in behavior between the two groups. Therefore, you cannot legitimately say that gender differences—rather than any of these other differences—caused your two groups to behave differently.

To help emphasize that you can make causal statements only about those independent variables that you randomly assign, randomly assigned variables are often called "true" independent variables or "strong" independent variables. In contrast, predictor variables that are not randomly assigned are called "weak" independent variables to highlight the fact that you can't determine whether they have an effect.

Reasons to Use Hybrid Designs

If you cannot make causal statements about the nonexperimental factor, why would you want to add a nonexperimental variable to your simple experiment? The most obvious and exciting reason is that you are interested in that nonexperimental variable.

To see how adding a nonexperimental variable (age of participant, introvert–extrovert, etc.) can spice up a simple experiment, consider the following simple experiment: Participants are either angered or not angered in a problem-solving task by a confederate who poses as another participant. Later, participants get an opportunity to punish or reward the confederate. Obviously, we would expect participants to punish the confederate more when they had been angered. This simple experiment, in itself, would not be very interesting.

TABLE **12.12**

The Hybrid Design: A Cross Between an Experiment and a Nonexperiment

	MEN	WOMEN	GENDER SIMPLE MAIN "EFFECTS"
Negative statements	10	12	2 (12 − 10 = 2)
Positive statements	8	14	6 (14 − 8 = 6)
Statement type simple main effects	−2 (8 − 10 = −2)	+2 (14 − 12 = +2)	

Averaging a factor's simple main effects gives us the factor's overall main effect:

Simple main effect of *Statement type* for men	−2
Simple main effect of *Statement type* for women	+2
Average effect (overall main effect) of *Statement type*	0/2 = 0
Simple main "effect" of **Gender** in the positive statements condition	2
Simple main "effect" of **Gender** in the negative statements condition	6
Average "effect" (overall main effect) of **Gender**	8/2 = 4

Comparing a treatment's simple main effects tells us whether there is an interaction:

Because there are differences between statement type's two simple main effects (i.e., −2 is different from +2), there is an interaction. In other words, because the effect of statement type is different for men than for women, there is a statement type × gender interaction

Note that the hybrid 2 × 2 design answers two questions that the simple experiment does not:

1. Do male and female participants differ on the dependent variable? (Answered by the gender main effect.)
2. Does the effect of statement type differ depending on which group (men or women) we are examining? (Answered by the gender × treatment interaction.)

Holmes and Will (1985) added a nonexperimental factor to this study—whether participants were Type A or Type B personalities. (People with Type A personalities are thought to be tense, hostile, and aggressive, whereas people with Type B personalities are thought to be more relaxed and less aggressive.) The results of this study were intriguing: If participants had *not* been angered, Type A participants were more likely to punish the confederate than Type B participants. However, if participants had been angered, Type A and Type B participants behaved similarly (see Figure 12.8).

Likewise, Hill (1991) could have done a relatively uninteresting simple experiment. He could have determined whether research participants are more likely to want to talk to a stranger if that stranger is supposed to be "warm" than if the stranger supposedly lacks warmth. The finding that people prefer to affiliate with nice people would not have been startling.

Fortunately, Hill conducted a more interesting study by adding another variable: need for affiliation. He found that participants who were high in need for affiliation were very likely to want to interact with an allegedly warm stranger, but very unlikely to want to interact with a stranger who

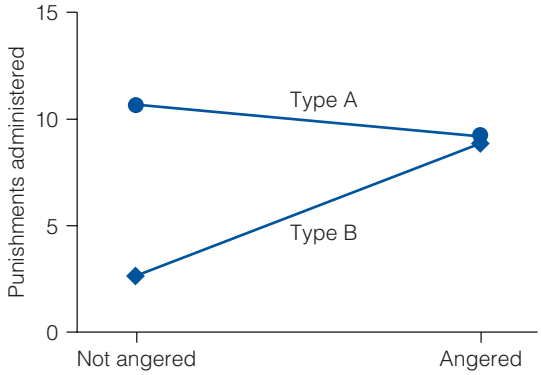

FIGURE **12.8** The Effect of Being Angered on the Aggressiveness of Type A and Type B Personality Types

Source: From Holmes, D. S., & Will, M. J. (1985). Expression of interpersonal aggression by angered and nonangered persons with Type A and Type B behavior patterns, by D. S. Holmes and M. J. Will, 1985, Journal of Personality and Social Psychology, 48, 723–727.

allegedly lacked warmth. For low need for affiliation participants, on the other hand, the alleged warmth of the stranger made little difference.

As you have seen, adding a nonexperimental factor can make a study more interesting. As you will see in the next sections, you can add a nonexperimental variable to a simple experiment for most of the same reasons you would add an experimental variable: to increase the generalizability of the findings, to look for a similarity effect, and to look for a moderating factor. In addition, you may add a nonexperimental factor to increase your chances of finding a significant effect for your experimental factor.

Increasing Generalizability

You could increase the generalizability of a simple experiment that used only men as participants by (a) using both men and women as participants and then (b) making gender of the participant a factor in your design. This design would allow you to determine whether the effect held for both men and women. For example, researchers (Crusco & Wetzel, 1984) wondered whether restaurant servers' "Midas touch"—touching customers results in bigger tips—holds for both men and women customers. (It does.) Some effects do not generalize across genders. For example, whereas men were *more* likely to say "yes" to a stranger's request to have sex than to say "yes" to a stranger's request to go on a date, women were *much less* likely to say "yes" to a stranger's request to have sex than to say "yes" to a stranger's request to go on a date (Clark & Hatfield, 2003).

In addition to seeing whether an effect generalizes across genders, you could see whether an effect generalizes across age, experience, or personality. For example, researchers have found that sleep-deprived younger drivers benefit more from a short nap than older drivers (Sagaspe et al., 2007); that both police officers and experienced judges are more likely to think that a videotaped confession is voluntary when the camera recording the confession is focused more on the suspect than on the detective (Lassiter, Diamond, Schmidt, & Elek, 2007); that, on math problems, people who normally do

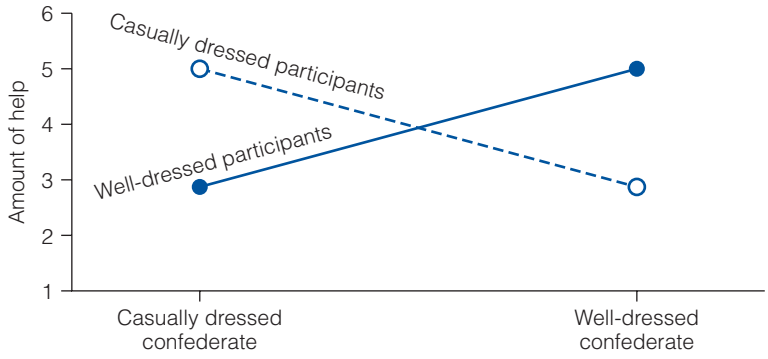

FIGURE **12.9** A Hybrid Design in Which an Interaction Represents a Similarity Effect

well in math are more likely to choke under pressure than people who normally do not do so well (Beilock & Carr, 2005); and that people with what could be described as aggressive personalities are just as affected by playing violent video games as other people (Anderson & Dill, 2000).

Studying Effects of Similarity: The Matched Factors Design

If you were interested in similarity, you might include some participant characteristic (gender, status, etc.) as a factor in your design, while manipulating the comparable (matching) experimenter or confederate factor. For example, if you were studying helping behavior, you could use style of dress of the participant (well-dressed/casual) and style of dress of the confederate as factors in your design. You might find this interaction: Well-dressed participants were more likely to help confederates who were well-dressed, but casually dressed participants were more likely to help confederates who were casually dressed. This interaction would suggest that similarity of dress influences helping behavior (see Figure 12.9).

Finding an Exception to the Rule: The Moderating Factors Design

Looking for the effects of similarity is not the only reason you would want to examine interactions involving participant characteristics. As we mentioned earlier, you might look at interactions involving participants to see whether a treatment that works with one type of person is as effective with another type of person. The treatment could be any intervention—from a therapy technique to a teaching style.

For instance, if you thought that intelligence would be a moderating variable for the effectiveness of computerized instruction, you might use intelligence as a factor in your design. To do this, you would first give your participants an IQ test and then divide them into two groups (above-average intelligence and below-average intelligence). Next, you would randomly assign the high-intelligence group to condition so that half of them were in computerized instruction and half were in lecture instruction. You would do the same for the low-intelligence group.

This hybrid study might reveal some interesting findings. Suppose you found that computerized instruction substantially increases learning for low-IQ children but slightly decreases learning for high-IQ children. If you had done only a simple experiment, you might have found a significant positive effect for the new teaching technique. On that basis, you might have made the terrible mistake of recommending that computerized instruction be used to teach all children.

Boosting Power: The Blocked Design

Suppose you were solely interested in seeing whether instructional technique had an effect and you had no interest in either IQ or the interaction between IQ and instructional technique. Even then, you might still include IQ as a factor in your experiment. Specifically, before the study begins, you might divide your participants into two *blocks* (groups): the low-IQ block and the high-IQ block. Then, you would randomly assign each member of the high-IQ group to instruction condition, thereby ensuring that half of the high-IQ participants are assigned to the computerized instruction condition and half are assigned to the lecture condition. Next, you would randomly assign each member of the low-IQ block to instruction condition.

In other words, you would do exactly the same study that we just recommended you do if you were looking at IQ as a moderating factor. However, this study would be called a **blocked design**: a factorial design in which, to boost power, participants are first divided into groups (blocks) on a participant variable (e.g., low-IQ block and high-IQ block) that is highly correlated with the dependent measure, and then participants from each block are randomly assigned to experimental condition.

The difference between doing this blocked design and doing the moderating factors study we just described is not *what* you are doing, but *why* you are doing it. If you are using a blocked design, you do not care about your blocking variable, and you do not care about the interaction between your blocking variable and your treatment. You are using the blocking variable solely to boost your chances of finding a statistically significant effect for your treatment.

To understand how the blocking variable will increase your chances of finding the treatment's effect, you first have to understand that just like decreasing the amount of dust on a microscope's lens increases your chances of seeing differences between cells, decreasing error variance increases your chances of seeing differences between treatment conditions. Then, you have to understand that blocked designs reduce error variance.

To understand how blocked designs reduce error variance, realize what error variance is—variability that is not accounted for in your study. If you use a simple experiment, individual differences in IQ are not accounted for; consequently, any variations in scores due to individual differences in IQ contribute to error variance. If, on the other hand, you use a blocked design that blocks on IQ, you account for some of the variance due to individual differences in IQ, thereby reducing your error variance. In a sense, you use your blocking variable to soak up variance that would otherwise be error variance. By shrinking the error variance, you make your treatment's effect easier to spot.

CONCLUDING REMARKS

We hope that you understand how factorial designs can help you refine your existing research ideas and generate new research ideas. We know that understanding factorial designs, one of the most common research methods in psychology, will increase your ability to read, understand, and evaluate other people's research.

SUMMARY

1. Factorial experiments allow you to look at the effects of more than one independent variable at a time.
2. The simplest factorial experiment is the one that looks at the effects of only two levels of two independent variables: the 2 × 2 ("two by two") experiment.
3. In addition to allowing you to see the individual effects of two factors in one experiment, the 2 × 2 experiment allows you to see whether the factors' combined effects are different from the sum of their individual effects.
4. Whenever the effect of combining two independent variables is different from the sum of their individual effects, you have an interaction. In other words, an interaction occurs when one independent variable's effect depends on the level of a second (moderating) variable. For example, the independent variable may have one effect when the second factor is absent and a different effect when the second factor is present.
5. Interactions often indicate that a general rule does not always apply. For instance, a treatment × distraction interaction indicates that the treatment does not have the same effect on people who are being distracted as on people who are not being distracted.
6. Interactions can most easily be observed by graphing your data. If your two lines aren't parallel, you may have an interaction.
7. A significant interaction usually qualifies main effects. Thus, if you find a significant interaction, you can't talk about your main effects without referring to the interaction.
8. Sometimes, an interaction represents similarity. For instance, in a 2 (place of learning: basement or top floor) × 2 (place of testing: basement or top floor) factorial experiment, an interaction may reveal that it is best to be tested in the *same* place you learned the information.
9. The following summarizes the mathematics of an ANOVA summary table for a factorial design:

SOURCE OF VARIANCE (*SV*)	SUM OF SQUARES (*SS*)	DEGREES OF FREEDOM (*df*)	MEAN SQUARE (*MS*)	*F*
A	*SS A*	Levels of *A* − 1	*SSA/df A*	*MSA/MSE*
B	*SS B*	Levels of *B* − 1	*SSB/df B*	*MSB/MSE*
A × *B* Interaction	*SS (A × B)*	*df A* × *df B*	*SS/df A* × *B*	*MS (A × B)/MSE*
Error	*SSE*	Participants − Groups	*SSE/df E*	
Total	*SS A* + *SS B* + *SS(AXB)* + *SSE*	Participants − 1		

10. With the hybrid factorial design, you can look at an experimental factor and a factor that you do not manipulate (personality, gender, age) in the same study. However, because you did not manipulate the nonexperimental factor, you cannot say that you know anything about the effects of your nonexperimental factor.
11. Once you have an idea for a simple experiment, you can easily expand that idea into an idea for a factorial experiment. For example, you could add a replication factor (such as stimulus set) to try to establish the generalizability of your treatment's effect. In that case, you would not be expecting a significant interaction. Alternatively, if you wanted to show that the treatment didn't have the same effect under all circumstances, you could add a potential moderating variable. In that case, you would be expecting a significant interaction between the treatment and the factor that you believe will moderate its effect.
12. If you have a nonmanipulated factor (e.g., participant's age), you can look at differences between groups on this factor. However, even though these differences are called main effects of the factor, do not make the mistake of thinking that these differences represent effects of the factor.

KEY TERMS

factorial experiments (*p. 418*)
simple main effect (*p. 422*)
overall main effect (*p. 424*)
interaction (*p. 425*)
crossover (disordinal) interaction (*p. 441*)
systematic replication (*p. 451*)
stimulus sets (*p. 451*)
blocked design (*p. 458*)

EXERCISES

1. What is the difference between
   a. a simple main effect and an overall main effect?
   b. an overall main effect and an interaction?
2. Can you have an interaction without a main effect?
3. Suppose an experimenter looked at the status of speaker and rate of speech on attitude change and summarized the experiment's results in the following table. Describe the pattern of those results in terms of main effects and interactions. Assume that all differences are statistically significant.

	STATUS OF SPEAKER	
Rate of Speech	**Low Status**	**High Status**
Slow	10	15
Fast	20	30

Attitude Change

4. Describe the pattern of results in the following table in terms of main effects and interactions. Assume that all differences are statistically significant.

	STATUS OF SPEAKER	
Rate of Speech	**Low Status**	**High Status**
Slow	10	15
Fast	20	25

Attitude Change

5. Half the participants receive a placebo. The other half receive a drug that blocks the effect of endorphins (pain-relieving substances, similar to morphine, that are produced by the brain). Half the placebo group and half the drug group get acupuncture. Then, all participants are asked to rate the pain of various shocks on a 1-to-10 (*not at all painful* to *very painful*) scale. The results are as follows: placebo, no acupuncture

group, 7.2; placebo, acupuncture group, 3.3; drug, no acupuncture group, 7.2; drug and acupuncture group, 3.3.
   a. Graph the results.
   b. Describe the results in terms of main effects and interactions (making a table of the data may help).
   c. What conclusions would you draw?

6. The following table is an incomplete ANOVA summary table of a study looking at the effects of similarity and attractiveness on liking. Complete the table. (Hint: If you are having trouble, consult Box 12.2 or the sample ANOVA summary table in Summary point 9.) Then, answer these three questions.
   a. How many participants were used in the study?
   b. How many levels of similarity were used?
   c. How many levels of attractiveness were used?

SV	*SS*	*df*	*MS*	*F*
Similarity *(S)*	10	1	—	—
Attractiveness *(A)*	—	2	20	—
$S \times A$ interaction	400	—	200	—
Error	540	54	—	
Total	990	59		

7. A professor does a simple experiment. In that experiment, the professor finds that students who are given lecture notes do better than students who are not given lecture notes. Imagine that you are asked to replicate the professor's simple experiment as a 2×2 factorial.
   a. What variable would you add to change the simple experiment into a 2×2?
   b. Graph your predictions.
   c. Describe your predictions in terms of main effects and interactions.

8. A lab experiment on motivation yielded the following results:

GROUP	PRODUCTIVITY
No financial bonus, no encouragement	25%
No financial bonus, encouragement	90%
Financial bonus, no encouragement	90%
Financial bonus, encouragement	90%

   a. Make a 2×2 table of these data.
   b. Graph these data (for help with graphing, see Box 12.1).
   c. Describe the results in terms of main effects and interactions. Assume that all differences are statistically significant.
   d. Interpret the results.

9. A memory researcher looks at the effects of processing time and rehearsal strategy on memory.

GROUP	PERCENT CORRECT
Short exposure, simple strategy	20%
Short exposure, complex strategy	15%
Long exposure, simple strategy	25%
Long exposure, complex strategy	80%

   a. Graph these data.
   b. Describe the results in terms of main effects and interactions. Assume that all differences are statistically significant.
   c. Interpret the results.

10. Suppose a researcher wanted to know whether lecturing was more effective than group discussion for teaching basic facts. Therefore, the researcher did a study and obtained the following results:

SOURCE OF VARIANCE	*SS*	*df*	*MS*	*F*
Teaching (*T*)	10	1	10	5
Introversion/ Extroversion (*I*)	20	1	20	10
$T \times I$ interaction	50	1	50	25
Error	100	50	2	

a. What does the interaction seem to indicate?
b. Even if there had been no interaction between teaching and extroversion, would there be any value in including the introversion–extroversion variable? Explain.
c. What, if anything, can you conclude about the effects of introversion on learning?

WEB RESOURCES

1. Go to the Chapter 12 section of the book's student website and
   a. Look over the concept map of the key terms.
   b. Test yourself on the key terms.
   c. Take the Chapter 12 Practice Quiz.
2. Download the Chapter 12 tutorial to practice the following:
   a. interpreting ANOVA tables
   b. interpreting graphs of results of factorial experiments.
3. Do an ANOVA using a statistical calculator by going to the "Statistical Calculator" link.

CHAPTER 13

Matched Pairs, Within-Subjects, and Mixed Designs

The art of being wise is the art of knowing what to overlook.
—William James

CHAPTER OVERVIEW

In Chapters 10, 11, and 12, you learned that you could perform an internally valid experiment by independently and randomly assigning participants to groups. Although you understand the logic of randomly assigning participants to groups, you may still have two basic reservations about between-subjects designs.

First, you may believe that these designs are wasteful in terms of the number of participants they require. For example, in the simple experiment, each participant is either in the control group *or* in the experimental group. If each participant was in both the control group *and* the experimental group, one participant could do the job of two.

Second, you may be concerned that between-subject designs are not powerful enough. You may believe that between-subject differences could hide treatment effects that would be detected if each participant acted as his or her own control. To illustrate, suppose you use a simple experiment to examine the effect of a video game "The Sims" on cooperation. If the effect of playing "The Sims" is small, then random differences between your groups could hide this effect. For example, suppose random assignment resulted in a comparison group that was naturally much more cooperative than the Sims group. In that case, if the Sims game slightly increased the Sims group's cooperation scores, the comparison group would still score higher on cooperation than the Sims group. Even if playing the game caused the Sims group to score slightly higher on cooperation than the comparison group, this difference may not be recognized as a treatment effect: In many cases, statistical tests could not rule out the possibility that such a small difference could be due to random differences between the two groups. If, on the other hand, you use each participant as his or her own control, the difference that the treatment created might be detected and found to be statistically significant.

You are rightfully concerned about the twin weaknesses of between-subjects experiments: They require many participants and have relatively little power to detect treatment effects. In this chapter, you will learn about designs that address these twin weaknesses: the matched-pairs design (a special type of between-subjects design) and two types of within-subjects

designs (sometimes called a "repeated-measures design"): the randomized within-subject design and the counterbalanced within-subjects design.

In the matched-pairs design, you first reduce between-subject differences by matching pairs of participants on a key characteristic (e.g., in a study of video game's effect on aggression, you might match participants on their scores on an aggression test). Then, you let random assignment and statistics take care of the effects of the remaining differences between participants.

In the randomized within-subjects design, you avoid the problem of between-subject differences by using participants as their own controls (e.g., you would compare each participant's score on the aggression measure after playing a violent video game with that same participants' score after playing a nonviolent videogame). Then, you let randomization take care of the effects of the remaining uncontrolled variables. By limiting the variables that randomization has to account for, the pure within-subjects design often has impressive power. For all its power, however, the randomized within-subjects design has some serious weaknesses. To build on its power but avoid those weaknesses, many researchers use what they consider a refinement of the randomized within-subjects design—the counterbalanced within-subjects design.

After learning about the two main types of pure within-subjects designs, you will learn about **mixed designs**: designs in which at least one factor is a within-subjects factor, and at least one factor is a between-subjects factor. In mixed designs, all participants get all levels of the within-subjects factor(s), but different participants get different levels of the between-subjects factor(s). For example, you might use a mixed design in which all participants played both the violent video game and the nonviolent video game, but some participants played the games in a hot room whereas others played the game in a normal temperature room. Mixed designs are popular because they can combine the power of a within-subjects design with the strengths of a between-subjects design.

Finally, you will learn how to weigh the trade-offs involved in choosing among various experimental designs. Thus, by the end of this chapter, you will be better able to choose the best experimental design for your research problem.

THE MATCHED-PAIRS DESIGN

If you do not have enough participants to do a powerful simple experiment, you might use a design, such as a matched-pairs design, that requires fewer participants. As you will see, the **matched-pairs design** combines the best aspects of matching and random assignment: It uses matching to reduce the effects of irrelevant variables, and it uses random assignment to establish internal validity.

Procedure

In the matched-pairs design, you first measure your participants on a variable that correlates with the dependent measure. For example, if you were doing a memory experiment, you might first give all your participants a memory test. Next, you would rank their scores on this memory test from lowest to highest. Then, you would pair the two highest scorers, the next two highest scorers, and so on. This would give you pairs of participants with similar scores on the memory pretest. Finally, you would randomly assign one member of each pair to the control group and the other member to the experimental group (e.g., you might assign random numbers to all the participants and then put the member of the pair with the higher random number in the experimental condition and the lower-scoring member in the control condition).

Considerations in Using Matched-Pairs Designs

You now have a general idea of how to conduct a matched-pairs experiment. You also know how it compares to a simple experiment: Unlike a simple experiment, it uses matching; like a simple experiment, it uses random assignment (see Table 13.1). But should you use a matched-pairs experiment instead of a simple experiment? When considering a matched-pairs design, you ask four questions:

1. Can you find an effective matching variable?
2. Will matching give you more power?
3. Will matching harm external validity?
4. Will matching harm construct validity?

Finding an Effective Matching Variable

As we suggested earlier, you can make effective use of the matched-pairs design only if you can create pairs that are very similar to each other in

TABLE **13.1**
Comparing the Matched Design with the Simple Experiment

MATCHED DESIGN	SIMPLE EXPERIMENT
First, *match* participants on key characteristics.	*No matching.*
Then, *randomly assign* each member of the pair to condition.	*Randomly assign* participants to condition.

terms of the dependent measure. The most direct way to get such pairs is to start your study by giving all the participants the dependent measure as a pretest and then matching participants based on their pretest scores. Thus, in a memory experiment, participants could be matched based on scores on an earlier memory test; in a maze-running experiment, participants could be matched based on scores on an earlier maze-running trial.

If you cannot match on pretest scores, you may have to search the research literature (see Web Appendix B) to find a matching variable. If you are lucky, you will find matching variables that other researchers have used. More likely, however, you will find out what variables correlate with your dependent measure. Unfortunately, after doing your library research, you may find that (a) there are no variables that have a strong, documented relationship with performance on the dependent measure, or that (b) there are good matching variables, but for ethical or practical reasons you cannot use them.

Power

You want to find an appropriate matching variable so that your study will have adequate **power**: the ability to find differences between conditions. Indeed, the reason you may choose a matched-pairs design is to avoid the power problems that plague researchers who use other types of between-subject designs.

As we discussed in Chapter 10, researchers who rely exclusively on random assignment to make groups similar lose power because individual differences between participants hide treatment effects. Specifically, because participants differ from each other, between-subjects researchers can't assume that the treatment group and the no-treatment group are extremely similar before the start of the experiment—especially if the groups are small. Consequently, if the groups differ at the end of the experiment, these researchers may not know whether this difference is due to the treatment or to the groups being different before the experiment began. Indeed, if a simple experiment has fewer than 30 participants, even a large difference between the treatment and no-treatment groups could be entirely due to random error.

If matching makes your groups extremely similar to each other before the experiment begins, then there isn't much random error due to individual differences to hide your treatment effects. Therefore, the same, small difference that would not be statistically significant with a simple experiment may be significant with a matched-pairs design.

How can a matched-pairs design give you more power than a simple experiment? The key, as we mentioned before, is that the matched-pairs design reduces random error, allowing the treatment effect to be seen as statistically significant. Mathematically, the matched-pairs design is more likely to find a statistically significant treatment effect because (a) the reduced random error results in larger t values and (b) larger t values are more likely to be statistically significant.

Why would the t value be larger in a matched-pairs design? Recall that the t value equals the difference between the means of the two conditions *divided by* an estimate of *random error* (the standard error of the difference). So, *with less random error,* the difference between groups is divided by less, and so the *t value becomes larger* (and thus more likely to be statistically significant). For example, if the standard error of the difference for a simple

experiment is 6 seconds, then a difference of 6 seconds between conditions would yield a *t* value of 1.0 (because 6/6 = 1.0)—a *t* value too low to be statistically significant. However, if a matched-pairs design reduced random error so much that the standard error of the difference was only 1.0, then that same difference of 6 seconds would yield a *t* value of 6.0 (because 6/1 = 6.0)—a *t* value that would probably be statistically significant. In other words, if matching limits the effects of individual differences, you may be able to find relatively small treatment effects.

But what if matching fails to reduce random error? For example, suppose a researcher matched participants on shoe size. In that case, the *t* value will be roughly the same as it would have been in the simple experiment because matching hasn't reduced the amount of random error in the study. In that case, the matched-pairs design would then be *less powerful* than the simple experiment.

To understand why poor matching leads to a matched-pairs design that is less powerful than a simple experiment, you need to know two facts: (1) Matched-pairs designs have half the degrees of freedom of a same-sized simple experiment, and (2) all other things being equal, fewer degrees of freedom means less power. We'll now take a closer look at these two facts.

By using a matched-pairs design instead of a simple experiment, you lose half your degrees of freedom because, whereas degrees of freedom for a simple experiment equals number of *participants*−2, the degrees of freedom for a matched-pairs study equals number of *pairs*−1. Thus, if you used 20 participants in a simple experiment, you would have 18 degrees of freedom (two fewer than the number of participants). But if you used 20 participants (10 pairs) in a matched-pairs design, you would have only 9 degrees of freedom (one fewer than the number of pairs).

Critical Values of *t*

LEVEL OF SIGNIFICANCE FOR TWO-TAILED *t* TEST	
df	.05
1	12.706
9	2.262
18	2.101
60	2.000
120	1.980

Losing degrees of freedom can cause you to lose power. As you can see by looking at this mini *t* table, the fewer degrees of freedom you have, the larger your *t* value must be to reach significance. For example, with 18 degrees of freedom (what you'd have if you tested 20 participants in a simple experiment), you would need only a *t* value of 2.101 for your results to be statistically significant at the .05 level. On the other hand, with 9 degrees of freedom (what you'd have if you tested 20 participants [10 pairs of participants] in a matched-pairs experiment), your *t* value

would have to be at least 2.262 to be statistically significant at the .05 level. That is, a difference between your treatment conditions that would have been big enough to be statistically significant if you had used a simple experiment might not be statistically significant with a matched-pairs design in which you matched on a variable that did not correlate with your measure. Thus, if you obtain the same *t* value with the matched-pairs design as you would have obtained with a simple experiment, the matched-pairs design costs you power.

If your matching is any good, however, you should not get the same *t* value with a matched-pairs design as with a simple experiment. Instead, you will get a larger *t* value with a matched-pairs design because you have reduced a factor that shrinks *t* values—random error due to differences between participants. Usually, the increase in the size of the *t* value will more than compensate for the degrees of freedom you will lose. Thus, as long as you can match participants on a relevant variable, you will get more power by switching from a simple experiment to a matched-pairs design.

External Validity

Power is not the only consideration in deciding to use a matched-pairs design. You may use—or avoid—matching for reasons of external validity.

Matched-Pairs Designs May Have Good External Validity. A matched-pairs design may have more external validity than an equally powerful simple experiment. Why? Because unlike the simple experiment, the matched-pairs design can have power without limiting who can be in the experiment.

To obtain adequate power, a researcher using a simple experiment may have to severely restrict the kind of individual who can be in the study. That is, to reduce the degree to which differences between participants create random differences between treatment and no-treatment groups, the experimenter may be forced to use participants who are all very similar. For example, to create a simple experiment that would be as powerful as a matched-pairs design, an experimenter might need to limit participants to male, albino rats between 180 and 185 days of age. Another researcher might attempt to reduce random error due to individual differences by allowing only middle-class women with IQs between 115 and 120 to be in the experiment.

With a matched-pairs design, however, you can reduce random differences between the treatment and no-treatment groups without choosing participants who are all alike. Because you can reduce random error by matching up the participants you do have rather than by limiting the kinds of participants you can have, the matched-pairs design may allow you to generalize your results to a broader population.

Matched-Pairs Designs May Have Poor External Validity. Matched-pairs designs, however, do not always have better external validity than simple experiments. For example, if participants drop out of the study between the time they are tested on the matching variable and the time they are to perform the experiment, matching will reduce the generalizability of your results. For instance, suppose you start off with 16 matched pairs, but end up with only 10 pairs. In that case, your experiment's external validity is compromised

because your results may not apply to individuals resembling the participants who dropped out of your experiment.

Even if participants do not drop out, matching may still harm external validity because your results generalize only to situations in which individuals perform the matching task before getting the treatment. To illustrate, imagine that an experimenter uses a matched-pairs design to examine the effect of caffeine on anxiety. In that experiment, participants take an anxiety test, then either consume caffeine (the experimental group) or do not (the control group), and then take the anxiety test again. Suppose that the participants receiving caffeine become more anxious than those not receiving caffeine.

Can the investigator generalize her results to people who have not taken an anxiety test before consuming caffeine? No, it may be that caffeine increases anxiety only when it is consumed after taking an anxiety test. For example, taking the anxiety test may make participants so concerned about their level of anxiety that they interpret any increase in arousal as an increase in anxiety. Because of the anxiety test, the arousal produced by caffeine—which might ordinarily be interpreted as invigorating—is interpreted as anxiety.

Construct Validity

In the caffeine study we just discussed, taking the anxiety test before and after the treatment might make participants aware that the experimenter is looking at the effects of a drug on anxiety. The participants' awareness of the hypothesis may harm the study's construct validity. For example, if participants believe that the hypothesis is that the drug will increase anxiety, they may act more anxious to help the researcher prove the hypothesis.

However, the fact that participants guess the hypothesis does not, by itself, ruin the experiment's construct validity. For instance, if you used a treatment condition and a placebo condition, it does not matter whether participants think that taking a pill is supposed to increase anxiety. Because both groups have the same hypothesis ("The pill I took will increase my anxiety"), knowing the hypothesis would not cause the treatment group to differ from the placebo group. Therefore, a significant difference between groups would have to be due to the treatment (the drug in the treatment group's pill).

If, on the other hand, your independent variable manipulation has poor construct validity, matching will make your manipulation's weaknesses more damaging. To see how matching can magnify a manipulation's weaknesses, imagine that the caffeine study used an empty control group (nothing was given to the participants who did not receive the treatment). The experimental group participants fill out an anxiety measure, take a pill, and then fill out another anxiety measure. The experimental group participants might think that the pill is supposed to increase their anxiety level, thereby causing them to be more anxious—or at least, to report being more anxious. The control group participants, not having been given a pill, would not expect to become more anxious. Consequently, a significant difference between the groups might be due to the two groups acting on different beliefs about what the researchers expected, rather than to any ingredient in the pill.

Analysis of Data

We have talked about how matching, by making your study powerful, can help you obtain a significant difference. We have also warned you about external validity and construct validity problems that should make you cautious when interpreting such significant differences. But how do you know whether you have a significant difference?

As we have already suggested, you should *not* use a regular, between-subjects *t* test. That test compares the overall, average score of the treatment group with the overall, average score of the no-treatment group.

With a matched-pairs design, you need a test that will allow you to compare the score of one member of a matched pair directly with the score of the other member—and to make that comparison for each of your pairs. If you have ratio or interval scale data,[1] you can make those comparisons using the **dependent groups *t* test.**[2] If you plan to do a dependent groups *t* test by hand, see Appendix E. If you plan to have a computer do a dependent groups *t* test for you, see Box 13.1.

BOX 13.1 Using the Computer to Conduct a Dependent Groups *t* Test

When looking for a computer program to do an analysis on a matched-pairs design or on a two-condition within-subjects design, realize that the test you are using may go by at least five names: (a) *t* test for correlated samples, (b) *t* test for dependent samples, (d) *t* test for paired samples, (d) repeated-measures *t* test, and (d) within-subjects *t*. Realize also that you are not limited to using a *t* test. For example, you could compute a within-subjects analysis of variance (see Box 13.2).

If you use a *t* test, the computer should provide you with at least three sets of information. First, it should tell you the number of observations you had in each condition. Thus, if you had four scores for condition 1, it should tell you that "*n*" for condition 1 was 4. Second, it should give you the mean (*M*) and the standard deviation (*SD*) for each condition. Third, it should give you the *t* value, the degrees of freedom (*df*) for the test, and the two-tailed probability (*p*) of obtaining a difference that great or greater between your two means if the null hypothesis were true. For example, a printout might look like the following two tables.

	CONDITION 1	CONDITION 2
n	4	4
M	6.25	2.5
SD	0.95	1.29
t	*df*	two-tailed *p*
15	3	.0006

You might report such results as follows.[a] "As predicted, significantly more words were recalled in the treatment condition ($M = 6.25$, $SD = 0.95$) than in the control condition ($M = 2.5$, $SD = 1.29$), $t(3) = 15.0$, $p < .05$."

[a] *M* stands for mean, *SD* stands for standard deviation (a measure of the variability of the scores), and p stands for the probability of obtaining a difference between conditions at least that large if the treatment had no effect. *SD* will usually be calculated as part of computing *t* (for more about *SD*, see Appendix E).

[1] If you have only ordinal data, you should use the sign test. If you don't know what type of data you have, consult Chapter 5.

[2] You can also analyze such data using a within-subjects ANOVA (see Box 13.2).

BOX 13.2 Using the Computer to Conduct a Within-Subjects Analysis of Variance

If you had conducted a matched-pairs study or a two-condition within-subjects study, you could analyze your data using a dependent groups *t* test (see Box 13.1) or a within-subjects ANOVA. If, instead of using the dependent groups *t* test as we did in Box 13.1, we used a within-subjects ANOVA, we would get similar results. For example, our printout might look like the following one.

DESCRIPTIVE STATISTICS

	CONDITION 1	CONDITION 2
n	4	4
M	6.25	2.5
SD	0.95	1.29

WITHIN-SUBJECTS ANOVA TABLE

SOURCE	*SS*	*df*	*MS*	*F*	*p*
Treatment	27.68	1	27.68	225	.0006
Error	0.37	3	0.123		

If you compare this ANOVA printout with the within-subjects printout in Box 13.1, you will note three similarities. First, the table listing the descriptive statistics in the within-subjects ANOVA printout is identical to the table listing the descriptive statistics in the within-subjects *t* test printout. The computer reports the same number of observations per condition, the same average for each condition, and the same variability of scores within each condition, regardless of whether you use a within-subjects *t* test or a within-subjects ANOVA.

Second, the *p* value for the treatment (.0006) in the within-subjects ANOVA table is the same as the p in the within-subjects *t* test. Both tests are equally likely to find a significant result.

Third, the *df* error (3) is the same as the *df* for the *t*. In both cases, *df equals number of participants minus two*.

Even the differences between the printouts reveal similarities. For example, the *F* value (225) is the *t* value (15) squared.

Given the similarities between the two types of analyses, you probably will not be surprised to learn that they would be written up similarly. Thus, you might report the above-described results as follows. "As predicted, significantly more words were recalled in the treatment condition ($M = 6.25$, $SD = 0.95$) than in the control condition ($M = 2.5$, $SD = 1.29$), $F(1,3) = 225.0$, $p < .05$."

If you had more than two levels of your independent variable, you could not use a within-subjects *t* test to analyze your data. You could, however, analyze such data with a multiple-level within-subjects ANOVA.

If you were to analyze such data with a multiple-level within-subjects ANOVA, your printout might resemble the printout of a two-level within-subjects ANOVA. Indeed, the most noticeable difference would be that your degrees of freedom will be different. For example, if you have 3 levels of the treatment, your treatment *df* will be 2.

As we have suggested, if you switch from looking at the printout of a two-level within-subjects design to looking at the printout of a three-level within-subjects design, you probably will not see a big difference. However, if you switch from looking at the printout from one computer program to another, you may notice a big difference. For example, in one program, a three-level, within-subjects ANOVA printout might look like the following printout.

WITHIN-SUBJECTS ANOVA TABLE

SOURCE	*SS*	*df*	*MS*	*F*	*p*
Treatment	12.133	2	6.067	26	.0001
Error	1.867	8	0.233		

However, the same analysis in another program might look like the table below—minus the footnotes. We added the footnotes to help you decipher the table.

TESTS OF WITHIN-SUBJECTS EFFECTS

	MEASURE				
SOURCE	TYPE III SUM OF SQUARES (*ss*)[a]	*df*	MEAN SQUARE[b]	*F*[c]	SIG.[d]
Treatment	12.133	2[e]	6.067	26	.000
Error (Treatment)	1.867	8	.233		

[a]Treat this column like the previous table's sum of squares (SS) column.
[b]Mean Square is calculated by dividing the Sum of Squares (12.133) by the df (2).
[c]$F = MS$ for the effect divided by MS error. The bigger F is, the more likely the results are to be statistically significant.
[d]This column represents how likely it is that one would obtain a result this large or larger if the null hypothesis were true. Traditionally, when the value in this column is less than .05, the results are considered "statistically significant."
[e]If there are 2 degrees of freedom (*df*), then there must be three levels of the "Treatment" variable.

BOX 13.2 Continued

In yet another program, the printout might look like the following table—minus the footnotes. (We added the footnotes to help you decipher the table.)

GENERAL LINEAR MODELS PROCEDURE REPEATED MEASURES ANALYSIS OF VARIANCE UNIVARIATE TESTS OF HYPOTHESES FOR WITHIN-SUBJECTS EFFECTS SOURCE: TREATMENT

df	TYPE III SUM OF SQUARES (SS)	MEAN SQUARE	*F VALUE*	P R > *F*[a]	GEISSER GREENHOUSE EPSILON PROB LEVEL[b] (G–T)	HUYNH FELDT EPSILON PROB LEVEL (H–F)
2	12.33	6.067	26	0.0001	0.0001	0.0001

[a] The value in this column corresponds to the p value or significance level that most programs give you.
[b] The probability value in this column or in the next column should be used if certain assumptions of the within-subjects ANOVA have been violated.

TABLE **13.2**
Advantages and Disadvantages of Matching

ADVANTAGES	DISADVANTAGES
More power because matching reduces the effects of differences between participants.	Matching makes more work for the researcher.
Power is not bought at the cost of restricting the subject population. Thus, results may, in some cases, be generalized to a wide variety of participants.	Matching may alert participants to the experimental hypothesis.
	Results cannot be generalized to participants who drop out after the matching task.
	The results may not apply to individuals who have not been exposed to the matching task prior to getting the treatment.

Conclusions About the Matched-Pairs Design

In summary, the matched-pairs design's weaknesses stem from matching (see Table 13.2). If you can't find an effective matching variable, matching may actually hurt power. If matching alerts participants to the purpose of your experiment, matching may hurt your construct validity. If participants drop

out of the experiment between the time they are measured on the matching variable and the time they are to be given the treatment, matching costs you the ability to generalize your results to the participants who dropped out. Finally, even if participants do not get suspicious and do not drop out, matching still costs you time and energy.

Although matching has its costs, matching usually offers one big advantage—power without restricting your subject population. Because the matched-pairs design combines the power of matching with the internal validity promoting properties of random assignment, the matched-pairs design is hard to beat when you can study only a few participants.

WITHIN-SUBJECTS (REPEATED MEASURES) DESIGNS

One set of designs that can beat the matched-pairs design, at least in terms of power, are the **within-subjects designs** (also called **repeated-measures designs**). In all within-subjects designs, each participant receives all the levels or types of the treatment that the experimenter administers, and the participant is measured after receiving each level or type of treatment. In the simplest case, each subject would receive only two levels of treatment: no treatment and the treatment. For example, a participant might complete the dependent-measure task (e.g., take an aggression test), get a treatment (e.g., play a violent video game), and repeat the dependent-measure task again (e.g., retake the aggression test). The experimenter would estimate the effect of the treatment by comparing how each participant scored when receiving the treatment (e.g., after playing a violent video game) with how that same participant scored when not receiving the treatment (e.g., before playing the violent video game).

Considerations in Using Within-Subjects Designs

You now have a general idea of how a within-subjects (repeated-measures) experiment differs from a between-subjects design (for a review, see Table 13.3).

TABLE **13.3**
Comparing Three Designs

	BETWEEN-SUBJECTS	MATCHED-PAIRS DESIGN	WITHIN-SUBJECTS
Role of random assignment	Randomly assign participants to treatment condition.	Randomly assign members of each pair to condition.	Randomly assign to sequence of treatment conditions.
Approach to dealing with the problem that differences between participants may cause differences between the treatment and no-treatment conditions.	Allow random assignment and statistics to account for any differences between conditions that could be due to individual differences.	Use matching to reduce the extent to which differences between conditions could be due to individual differences. Then, use random assignment and statistics to deal with the effects of individual differences that were not eliminated by matching.	Avoid the problem of individual differences causing differences between conditions by comparing each participant's performance in one condition with his or her performance in the other condition(s).

But what do you have to gain—or lose—by using a within-subjects design instead of a between-subjects design? As you'll soon see, by using a within-subjects design instead of a between-subjects design, you will gain power; however, you may lose internal validity.

Increased Power

Despite potential problems with the within-subjects design's internal validity, the within-subjects design is extremely popular because it increases power in two ways.

The first way is similar to how the matched-pairs design increases power—by reducing random error. As you may recall, the matched-pairs experimenter tries to reduce random error by reducing individual differences by comparing similar participants with one another. Within-subjects experimenters are even more ambitious: They want to eliminate random error due to individual differences. Therefore, they do not compare one participant with another participant; instead, they compare each participant's score under one condition with that same participant's score under another condition.

The second way the within-subjects design increases power is by increasing the number of observations you get from each participant. The more observations you have, the more random error will tend to balance out; the more random error balances out, the more power you will have. With between-subjects designs, the only way you can get more observations is to get more participants because you can only get one observation per participant. But in a within-subjects experiment, you get at least two scores out of each participant. In the simplest case, your participants serve double duty by being in both the control and experimental conditions. In more complex within-subjects experiments, your participants might do triple, quadruple, or even octuple duty. For example, in a study of how men's muscularity affected women's ratings of men, Frederick and Haselton (2007) had participants do octuple duty. Specifically, to test their hypothesis that muscularity—up to a point—would increase attractiveness ratings, Frederick and Haselton had women rate the attractiveness of eight drawings that varied in muscularity. If Frederick and Haselton had used a purely between-subjects design, each participant would have made only one rating. However, because they used a within-subjects design, each participant could rate all eight figures.

Order Effects May Harm Internal Validity

As you intuitively realize, the main advantage of within-subjects designs is their impressive power. By comparing each participant with him or herself, even subtle treatment effects may be statistically significant.

However, as you may also intuitively realize, the problem with comparing participants with themselves is that, even without the treatment, participants may change over time. Consequently, the **order** (first or last) in which an event occurs within a sequence of events can be very important. For example, the lecture that might have been fascinating had it been the first lecture you heard that day might be only tolerable if it is your fourth class of the day. Because order affects responses, if a participant reacts differently to the first treatment than to the last, we have a dilemma: Do we have a treatment effect or an order effect?

TABLE **13.4**
In a Within-Subjects Design, the Treatment May Not Be the Only Factor Being Manipulated

	EVENTS THAT OCCUR BEFORE BEING TESTED	
	Drug 1 Condition	**Drug 2 Condition**
Between-Subjects Experiment	Get Drug 1	Get Drug 2
Within-Subjects Design	Get Drug 1	Get Drug 1
		Play Video Game
		Get Drug 2

To get a better idea of how **order (trial) effects** can complicate within-subjects experiments, let's examine a within-subjects experiment. Imagine being a participant in a within-subjects experiment where you take a drug (e.g., caffeine), play a video game, take a second drug (e.g., aspirin), and play the video game again.

If you perform differently on the video game the second time around, can the experimenters say that the second drug has a different effect than the first drug? No. The experimenters can't safely make conclusions about the difference between the two drugs because they are comparing your performance on trial 1, when you had been exposed to only one treatment (drug 1), to your performance on trial 2, by which time you had been exposed to three "treatments": (1) drug 1, (2) playing the game, and (3) drug 2 (see Table 13.4).

Four Sources of Order Effects

In the next few sections, you will see how being exposed to "treatments" other than the second drug can hurt the study's internal validity. We will start by showing you how the variable of *order* (first trial vs. second trial) may affect your performance. Specifically, we will look at four nontreatment reasons why you may perform differently on the task after the second treatment:

1. You may do better after the second treatment because you are performing the dependent-measure task a second time. For example, the practice you got playing the game after the first drug may help you when you play the game again.
2. You may do worse after the second treatment because you are bored with the dependent-measure task.
3. You may score differently because you are experiencing some delayed or lingering effects of the first treatment.
4. You may have figured out the experimental hypothesis right after you received the second treatment.

In summary, you need to be aware that the order in which participants get a treatment may affect the results. Thus, Treatment A may appear to have one kind of effect when it comes first, but may *appear* to have a different kind of effect when it comes second.

Practice Effects

If you perform better after the second treatment than you did after the first treatment, your improvement may merely reflect **practice effects**: You may have learned from the first trial. The first trial, in effect, trained you how to play the video game—although that wasn't the researcher's plan. Not surprisingly, practice effects are common: Participants often perform better as they warm up to the experimental environment and get accustomed to the experimental task. Unfortunately, rather than seeing that you improved because of practice, the researcher may mistakenly believe that you improved due to the treatment.

Fatigue Effects

If your performance is not enhanced by practice, it may decline due to **fatigue effects**.[3] You may do worse on later trials merely because you are becoming tired or less enthusiastic as the experiment goes on. Unfortunately, a researcher might interpret your fatigue as a treatment effect.

Treatment Carryover Effects

Practice and fatigue effects have nothing to do with any of the treatments participants receive. Often, practice and fatigue effects are simply due to getting more exposure to the dependent-measure task. Thus, in the video game example, performance may improve as you learn the game or worsen as you get bored with the game. However, exposure to the dependent measure is not the only thing that can affect performance in later trials. The effects of a treatment received before the first trial may affect responses in later trials. The effects of an earlier treatment on responses in later trials are called **carryover (treatment carryover) effects**.

To imagine treatment carryover effects, suppose that on Trial 1, the researcher gave you a tranquilizer and then measured your video game performance. On Trial 2, the researcher gave you an antidepressant and measured your video game performance. On Trial 3, the researcher gave you a placebo and measured your video game performance. If your performance was worst in the placebo (no-drug) condition, the researcher might think that your better performance on earlier trials was due to the drugs improving your performance. The researcher, however, could be wrong. Your poor performance in the placebo condition may be due to carryover effects from the previous treatments: You may just be starting to feel certain effects of the drugs that you consumed during the earlier trials. Depending on the time between the trials, you may be feeling either "high" or hung-over.

Sensitization Effects

A fourth factor that might cause you to perform differently after the second treatment is **sensitization**. Sensitization occurs when, after getting several different treatments and performing the dependent variable task several times, participants realize (become sensitive to) what the independent and dependent variables are, and thus, during the latter parts of the experiment, guess

[3] Fatigue effects could be viewed as cases in which performance is hurt by practice, whereas practice effects could be viewed as cases in which performance is improved by practice.

the experimental hypothesis and play along with it. For example, by the third trial of the video game experiment, you should realize that the experiment had something to do with the effects of drugs on video game performance.

Note that sensitization has two effects. First, it threatens construct validity because participants figure out what the hypothesis is and thus may be acting to support the hypothesis rather than reacting to the treatment. Second, it threatens internal validity because it makes participants behave differently during the last trial (when they know the hypothesis) than they did during the first trial (when they did not know the hypothesis).

Review of the Four Sources of Order Effects

You have seen that because of practice, fatigue, carryover, and sensitization, the sequence in which participants receive the treatments could affect the results. For example, suppose participants all received the treatments in this sequence: Treatment A first, Treatment B second, and Treatment C last. Even if none of the treatments had an effect, the effect of order (first vs. second vs. last) might make it look like the treatments had different effects.

If practice effects caused participants to do better on the last trial, participants would do best on the trial where they received Treatment C. Thus, even if none of the treatments had an effect, the investigator might mistakenly believe that Treatment C improves performance.

If, on the other hand, fatigue effects caused participants to perform the worst on the last treatment condition, participants would do worst on the trial where they received Treatment C. Thus, even if none of the treatments had an effect, the investigator might mistakenly believe that Treatment C decreases performance.

Treatment carryover effects might also affect performance on the last trial. For example, if the effect of Treatment B is helpful but delayed, it might help performance on the last trial. If, on the other hand, the effect of Treatment B is harmful but delayed, it might harm performance on the last trial. Thus, even if Treatment C has no effect, the investigator might mistakenly believe that Treatment C is harmful (if Treatment B's delayed effect is harmful) or that Treatment C is helpful (if Treatment B's delayed effect is helpful).

Sensitization might also create the illusion that Treatment C has an effect. The participants were most naïve about the experimental hypothesis when receiving the first treatment (Treatment A), least naïve when receiving the last treatment (Treatment C). Thus, the ability of the participant to play along with the hypothesis increased as the study went on. Changes in the ability to play along with the hypothesis may create order effects that could masquerade as treatment effects.

Dealing With Order Effects

You have seen that (a) the sources of order effects are practice, fatigue, carryover, and sensitization; and that (b) order effects threaten the internal validity of a within-subjects design. How can you use this knowledge to prevent order effects from threatening your experiment's internal validity?

Minimizing Each of the Individual Sources of Order Effects

Perhaps the best place to start to reduce the effect of order is to attack the four root causes of order effects: practice, fatigue, carryover, and sensitization.

Minimizing Practice Effects. To minimize the effects of practice, you can give participants extensive practice before the experiment begins. For example, if you are studying maze running and you have the rats run the maze 100 times before you start administering treatments, they've probably learned as much from practice as they can. Therefore, it's unlikely that the rats will benefit greatly from the limited practice they get during the experiment.

Minimizing Fatigue Effects. You can reduce fatigue effects by making the experiment interesting, brief, and undemanding.

Minimizing Treatment Carryover Effects. You can reduce carryover effects by lengthening the time between treatments to allow adequate time for the effect of earlier treatments to wear off before the participant receives the next treatment. For instance, if you were looking at the effects of drugs on how well rats run a maze, you might reduce treatment carryover effects by spacing your treatments a week apart (for example, antidepressant pill, wait a week, anti-anxiety pill, wait a week, placebo).

Minimizing Sensitization Effects. You can reduce sensitization by preventing participants from noticing that you are varying anything (Greenwald, 1976). For example, suppose you were studying the effects of different levels of full-spectrum light on typing performance. In that case, there would be three ways that you could prevent sensitization.

First, you could use very similar levels of the treatment in all your conditions. By using slightly different amounts of full-spectrum light, participants may not realize that you are actually varying amount of light.

Second, you could change the level of the treatment so gradually that participants do not notice. For example, while you gave participants a short break in between trials, you could change the lighting level watt by watt until it reached the desired level.

Third, you might be able to reduce sensitization effects by using good placebo treatments. In this example, rather than using darkness as the control condition, you could use light from a normal bulb as the control condition.

A General Strategy for Reducing Order Effects

To this point, we have given you some strategies to reduce practice effects, to reduce fatigue effects, to reduce carryover effects, and to reduce sensitization (see Table 13.5 for a review). However, by reducing the number of experimental conditions, you can reduce all four causes of order effects at once because there will be fewer opportunities for them to affect your study.

To see how fewer conditions leads to fewer order-effect problems, compare a within-subjects experiment that has 11 conditions with one that has only 2 conditions. In the 11-condition experiment, participants have 10 opportunities to practice on the dependent-measure task before they get

TABLE **13.5**
Order Effects and How to Minimize Their Impact

EFFECT	EXAMPLE	WAYS TO REDUCE IMPACT
Practice Effects	Getting better on the task due to becoming more familiar with the task or with the research situation.	Give extensive practice and warm-up before introducing the treatment.
Fatigue Effects	Getting tired as the study wears on.	Keep study brief, interesting.
Carryover Effects	Effects of one treatment lingering and affecting responses on later trials.	Use few levels of treatment. Allow sufficient time between treatments for treatment effects to wear off.
Sensitization	As a result of getting many different levels of the independent variable, the participant—during the latter part of the study—becomes aware of what the treatment is and what the hypothesis is.	Use subtly different levels of the treatment. Gradually change treatment levels. Use few treatment levels.

the last treatment; in the 2-condition experiment, participants only have one opportunity for practice. The 11-condition participants have 11 conditions to fatigue them; 2-condition participants only have 2. In the 11-condition experiment, there are 10 treatments that could carry over to the last trial; in the 2-condition experiment there is only 1. Finally, in the 11-condition experiment, participants have 11 chances to figure out the hypothesis; in the 2-condition experiment, they only have 2 chances.

Mixing Up Sequences to Try to Balance Out Order Effects: Randomizing and Counterbalancing

Although you can take steps to reduce the impact of order, you can never be sure that you have eliminated its impact. Therefore, if you gave all your participants Treatment A first and Treatment B second, you could not be sure that the difference between the average of the Treatment A scores and the average of the Treatment B scores was due to a treatment effect. Instead, the difference could simply be due to an order (trials: first vs. second) effect.

To avoid confusing an order (trials) effect for a treatment effect, you should not give all your participants the same sequence of treatments. For example, in a two-condition study, you should not give all of your participants the treatments in this sequence: Treatment A first, Treatment B second. Instead, some participants should get the treatment sequence: Treatment B first and then Treatment A.

There are two basic approaches you could use to make sure that not all participants get the treatments in the same sequence: (1) Randomize the sequence of treatments for each participant, or (2) counterbalance the sequence of treatments.

RANDOMIZED WITHIN-SUBJECTS DESIGNS

You can mix up the sequences by randomly determining, for each participant, which treatment they get first, which treatment they get second, and so on. If you use this randomization strategy to sequence each participant's series of treatments, you have a randomized within-subjects design.

Procedure

The **randomized within-subjects design** is very similar to the matched-pairs design. For example, the procedural differences between the two-condition, randomized, within-subjects experiment and matched-pairs experiment stem from a single difference: In the within-subjects experiment, you get a pair of scores from a single participant, whereas in the matched-pairs design, you get a pair of scores from a matched pair of participants. Thus, in the matched-pairs case, each participant only gets one treatment, but in the within-subjects experiment, each participant gets two treatments.

Other than each participant receiving more than one treatment, the two designs are remarkably similar. The matched-pairs researcher randomly determines, for each pair, who will get what treatment. In some pairs, the first member will get Treatment A, whereas the second member will get Treatment B; in other pairs, the first member will get Treatment B, whereas the second member will get Treatment A.

The within-subjects researcher randomly determines, for each individual, the sequence of the treatments. For some individuals, the first treatment will be Treatment A (and the second treatment will be Treatment B); for other individuals, the first treatment will be Treatment B (and the second treatment will be Treatment A). In short, whereas the matched-pairs experimenter randomly assigns members of pairs to different treatments, the within-subjects experimenter randomly assigns individual participants to different sequences of treatments.

To see the similarities and differences between the matched-pairs and within-subjects designs, imagine that we are interested in whether observers' judgments about other people are influenced by irrelevant information. Specifically, we want to see whether pseudorelevant information (information that seems relevant but really isn't relevant) affects whether observers see others as passive or assertive. Therefore, we produce pseudorelevant descriptions ("Bill has a 3.2 GPA and is thinking about majoring in psychology") and "clearly irrelevant" descriptions ("Bob found 20 cents in a pay phone in the student union when he went to make a phone call").

In a matched-pairs design, you would match participants—probably based on how assertively they tend to rate people. Then, one member of the pair would read a "pseudorelevant" description while the other read a "clearly irrelevant" description. After reading the information, each participant would rate the assertiveness of the student he read about on a 9-point scale ranging from "very passive" to "very assertive."

In a randomized within-subjects design, on the other hand, each participant would read both "pseudorelevant" and "clearly irrelevant" descriptions. After reading the information, they would rate the assertiveness of each of these students on a 9-point scale ranging from "very passive" to "very assertive." Thus, each participant would provide data for both the "pseudorelevant" condition and the "clearly irrelevant" condition. The sequence of the descriptions would be randomized, with some sequences having the pseudorelevant description first and others having the clearly irrelevant description first.

Hilton and Fein (1989) conducted such a randomized within-subjects experiment and found that participants judged the students described by pseudorelevant information as more assertive than students described by clearly irrelevant information. Consequently, Hilton and Fein concluded that even irrelevant information affects our judgments about people.

Analysis of Data

To analyze data from the two-condition within-subjects design, you can use the same dependent groups *t* test that you used to analyze matched-pairs designs.[4] The only difference is that instead of comparing each member of the pair with the other member of that pair, you compare each participant with him or herself. Because the dependent groups *t* test can be used to analyze data from a within-subjects design, it can also be called the *within-subjects t test*.

You do not have to use a within-subjects *t* test. For example, instead of using a within-subjects *t* test (see Box 13.1), you could use a within-subjects analysis of variance (see Box 13.2).

Conclusions About Randomized Within-Subjects Designs

As you might expect from two designs that can be analyzed with the same technique, the randomized within-subjects design and the matched-pairs design are very similar. In terms of procedures, the only real difference is that the matched-pairs experimenter randomly assigns members of pairs to treatments, whereas the randomized within-subjects experimenter randomly assigns individual participants to sequences of treatments. Because both designs have impressive power, both should be seriously considered if participants are scarce.

The randomized within-subjects design, however, has some unique strengths and weaknesses stemming from the fact that it collects more than one observation per participant (see Table 13.6). Because it uses individual participants (rather than matched pairs) as their own controls, the randomized within-subjects design is more powerful than the matched-pairs design—and more useful when you want to generalize your results to real-life situations in which individuals get more than one "treatment." Thus, if you were studying the effects of political ads, you might use a within-subjects design because, in real life, a person is likely to be exposed to more than one political ad about a candidate (Greenwald, 1976).

---

[4] If you have more than two conditions, you cannot use a *t* test. Instead, you must use either within-subjects analysis of variance (ANOVA) or multivariate analysis of variance (MANOVA).

TABLE **13.6**
Comparing the Matched-Pairs Design With the Within-Subjects Design

MATCHED-PAIRS DESIGN	WITHIN-SUBJECTS DESIGN
Powerful.	More powerful.
Order effects are *not* a problem.	Order effects are a serious problem.
Uses random assignment to balance out differences between participants.	Uses randomization to balance out order effects.
Useful for assessing variables that vary between subjects in real life.	Useful for assessing variables that vary within subjects in real life.

Although there are benefits to collecting more than one observation per participant, having to contend with order effects (practice, fatigue, carryover, and sensitization) is a major drawback. As we have suggested, you can try to minimize order effects, and you can *hope* that randomization will balance out the sequence of your treatments so that each condition comes first about the same number of times as it comes last.

COUNTERBALANCED WITHIN-SUBJECTS DESIGNS

Instead of merely hoping that chance might balance out the sequence of your treatments, you could make sure by using a **counterbalanced within-subjects design**. In this design, as in all within-subjects designs, each participant gets more than one treatment. Unlike other within-subjects designs, however, participants are randomly assigned to systematically varying sequences of conditions in a way that ensures that *routine order effects* are balanced out.[5] Thus, if you were studying two levels (A and B) of a factor, the counterbalanced design would ensure that half your participants got Treatment A first and that half got Treatment B first. Now that you understand the main objective of counterbalancing, let's look at an example to see how counterbalancing achieves this goal.

Procedure

If you were to use a counterbalanced design to study a two-level factor, you would randomly assign half of your participants to receive Treatment A first and Treatment B second, whereas the other half would receive Treatment B first and Treatment A second. By randomly assigning your participants to these counterbalanced sequences, most order effects will be neutralized. For example, if participants tend to do better on the second trial, this will not

[5] In football, for example, teams change sides every quarter and this usually balances out the effects of wind. However, if the wind shifts in the fourth quarter, counterbalancing fails to balance out the effects of wind. Similarly, if basketball teams change sides at the end of every half (as in international rules), but a rim gets bent (or fixed) during halftime, counterbalancing has failed to balance out the effects of different baskets.

help Treatment A more than Treatment B because both occur in the second position equally often.

Advantages and Disadvantages of Counterbalancing

By using a counterbalanced design, you have not merely balanced out routine order effects, but you have also added another factor to your design—the between-subjects factor of counterbalancing sequence. Adding the factor of counterbalancing sequence has two disadvantages and several advantages.

Disadvantages of Adding a Counterbalancing Factor

A minor disadvantage is that your statistical analysis is now more complex. Rather than using the dependent (within-groups) *t* test, you now have to use a mixed analysis of variance. This would be a major disadvantage if you had to compute statistics by hand. However, because computers can do these analyses for you, this disadvantage really is minor.

The major disadvantage of adding the two-level between-subjects factor of counterbalancing sequence is that you now need more participants than you did when you were planning to use a pure within-subjects design. You need two groups of participants to determine whether the two-level between-subjects factor of counterbalanced sequence has an effect. One of those groups will receive the A–B sequence, the other will receive the B–A sequence. To have enough power to see whether the A–B sequence leads to higher average scores than the B–A sequence, you will need at least 30 participants in each group.[6] In the pure within-subjects design, on the other hand, we were not comparing one group against another. Thus, in a sense, by going from a pure within-subjects design to a counterbalanced design, you are going from having zero levels of a between-subjects factor to having two levels of a between-subjects factor. As you may recall from our discussion of multiple-group experiments (Chapter 11), the more levels of a between-subjects factor you have, the more participants you need.

As you go beyond two levels of the independent variable, the number of different possible sequences—and thus the levels of the between-subjects factor of counterbalancing—explodes. For example, if you have 3 levels, there are 6 possible sequences (ABC, BCA, CAB, ACB, BAC, CBA), and thus counterbalancing could be a 6-level factor; if you have 4 levels, counterbalancing could be a 24-level factor, and if you have 5 levels, counterbalancing could be a 120-level factor.

To avoid the problem of having too many levels of the between-subjects factor of counterbalancing, you have two options. First, if you can administer the same treatment more than once to the same participant, you can get by with a single sequence. Ideally, this sequence would involve all possible orders (e.g., ABC, BCA, CAB, ACB, BAC, CBA). However, if you could present each treatment only twice, a sequence in which you first present the treatments in one order (e.g., ABC) and then present them in the reverse order (e.g., CBA) offers some protection from order effects. Thus, if you had five treatments, you might present them in this sequence: ABCDEEDCBA.

---

[6] In most cases, 30 participants per group is too few. Usually, researchers should have at least 60 participants per group (Cohen, 1990).

Second, rather than randomly assigning participants to every possible sequence (as you would in complete counterbalancing), you randomly assign participants to as many sequences as you have levels of your independent variable. Thus, if you have four treatments (A, B, C, and D), you would randomly assign participants to four sequences.

The first key to this partial counterbalancing is to select a set of sequences that, like complete counterbalancing, has every condition occur in every position equally often (e.g., Treatment A occurs first just as often as it occurs second, third, and fourth—and what is true of Treatment A is true of all your treatments). The simplest way to do this is to have each condition appear *once* in each position. Thus, if you had four treatments, treatment A would appear first in one sequence, second in one sequence, third in one sequence, and fourth in one sequence—and the same would be true of treatments B, C, and D.

The second key to this partial counterbalancing is to have each condition come before every other condition just as many times as it comes after that condition (e.g., Treatment A comes before Treatment B twice and after Treatment B twice). To get such a set of sequences, you would use a Latin Square (see Box 13.3). Note, however, that even with partial counterbalancing, you need more participants than you would with a pure, randomized within-subjects design.

BOX 13.3 Latin Square Designs: The ABCs of Counterbalancing Complex Designs

You have seen an example of the simplest form of counterbalancing in which one group of participants gets Treatment A followed by Treatment B (AB) and a second group gets Treatment B followed by Treatment A (BA). This simple form of counterbalancing is called AB, BA counterbalancing. Note that even this simple form of counterbalancing accomplishes two goals.

First, it guarantees every condition occurs in every position, equally often. Thus, in AB, BA counterbalancing, A occurs in first half the time and last half the time. The same is true for B: For half the participants, B is the first treatment they receive; for the other half, B is the last treatment they receive.

Second, each condition precedes every other condition just as many times as it follows that condition. That is, in AB, BA counterbalancing, A precedes B once and follows B once. This symmetry is called *balance*.

Although achieving these two objectives of counterbalancing is easy with only two conditions, with more conditions, counterbalancing becomes more complex. For example, with four conditions (A, B, C, D) you would have four groups. To determine what order the groups will go through the conditions, you would consult the following 4 × 4 Latin Square:

	POSITION			
	1	**2**	**3**	**4**
Group 1	A	B	D	C
Group 2	B	C	A	D
Group 3	C	D	B	A
Group 4	D	A	C	B

(Continued)

BOX 13.3 Continued

In this 4 × 4 complete Latin Square, Treatment A occurs in all four positions (first, second, third, and fourth), as do Treatments, B, C, and D. In addition, the square has balance. As you can see from looking at the square, every letter precedes every other letter twice and follows every other letter twice. For example, if you just look at Treatments A and D, you see that A comes before D twice (in Groups 1 and 2) and follows D twice (in Groups 3 and 4).

Balance is relatively easy to achieve for 2, 4, 6, 8, or even 16 conditions. But, what if you have 3 conditions? Immediately you recognize that with a 3 × 3 Latin Square, A cannot precede B the same number of times as it follows B. Condition A can either precede B twice and follow it once or precede it once and follow it twice. Thus, with an uneven number of conditions, you cannot create a balanced Latin Square.

One approach to achieving balance when you have an uneven number of treatment levels is to add or subtract a level so you have an even number of levels. However, adding a level may greatly increase the number of sequences and groups you need. Subtracting a level, on the other hand, may cause you to lose vital information. Therefore, you may not wish to alter your study to obtain an even number of levels. Fortunately, you can achieve balance with an uneven number of treatment levels by using two Latin Squares.* For instance, consider the 3 × 3 squares below.

If you randomly assign subjects to six groups, as outlined above, you ensure balance. See for yourself that if you take any two conditions, one condition will precede the other three times and will be preceded by the other condition three times.

	SQUARE 1 POSITION				SQUARE 2 POSITION		
	1	2	3		1	2	3
Group 1	A	B	C	Group 4	C	B	A
Group 2	B	C	A	Group 5	A	C	B
Group 3	C	A	B	Group 6	B	A	C

*Another option is to use incomplete Latin Square designs. However, the discussion of incomplete Latin Square designs is beyond the scope of this book.

Advantages of Adding a Counterbalancing Factor

The disadvantage of needing more participants is sometimes offset by being able to discover more effects. With the two-condition within-subjects experiment, you can obtain only one main effect (the treatment main effect). By adding the two-level factor of counterbalancing sequence, you converted the two-condition experiment into a 2 (the within-subjects factor of treatment) × 2 (the between-subjects factor of counterbalancing sequence) experiment, thus giving you more information. Specifically, you can look for two main effects and an interaction (see Table 13.7). By looking at these three effects, you can find out three things.

First, as was the case with the pure within-subjects design, you can find out whether the treatment had an effect by looking at the treatment main effect. In the experiment described in Table 13.7, you can look at the treatment main effect to find out whether forming images of words is a more effective memory strategy than making sentences out of the words.

Second, by looking at the counterbalancing sequence main effect, you find out whether the group of participants getting one sequence of treatments (A–B) did better than the participants getting the other (B–A) sequence. In the experiment described in Table 13.7, the question is, "Did Group 1 (who

TABLE **13.7**
A 2 × 2 Counterbalanced Design

The members of the first group get a list of words, are asked to form images of these words, and are asked to recall these words. Then, they get a second list of words, are asked to form a sentence with these words, and are asked to recall the words.

The members of the second group get a list of words, are asked to form a sentence with these words, and are asked to recall these words. Then, they get a second list of words, are asked to form images of those words, and are asked to recall those words.

GROUP 1	
First Task	**Second Task**
Form Images	Form Sentences
GROUP 2	
First Task	**Second Task**
Form Sentences	Form Images

Questions this study can address include the following:

1. Do people recall more when asked to form sentences than when asked to form images?
2. Do Group 1 participants recall more words than Group 2 participants? In other words, is one sequence of using the two different memory strategies better than the other?
3. Do people do better on the first list of words they see than on the second? That is, does practice help or hurt?

formed images first and then formed sentences) recall more words than Group 2 (who formed sentences first and then formed images)?"

Third, by looking at the treatment × counterbalancing interaction, you find out whether participants score differently on their first trial than on their second. Looking at the treatment × counterbalancing interaction allows you to detect what some people call a "trials effect" and what others call an "order effect."

But how can looking at an interaction tell you that participants score differently on the first trial than the second? After all, significant interactions usually indicate exceptions to general rules rather than indicating a general rule such as, "participants do better on the first trial."

The first step to seeing why a significant treatment × counterbalancing interaction tells you that participants score differently on the first trial than on the second is to imagine such an interaction. Suppose that participants who get Treatment A *first* score highest after receiving Treatment A, *but* participants who get Treatment B *first* score highest after receiving Treatment B. At one level, this is an interaction: The rule that participants score highest when receiving Treatment A only holds when they receive Treatment A first. However, the cause of this interaction is an order (trials) effect: Participants score highest on the first trial.

To get a clearer idea of what a counterbalanced study can tell us, let's look at data from the memory experiment we mentioned earlier. In that experiment, each participant learned one list of words by making a sentence out of the list and learned one list of words by forming mental images. Thus, like a within-subjects design, each participant's performance under one treatment condition (sentences) was compared with that same participant's performance under another treatment condition (images).

Like a two-group between-subjects design, participants were randomly assigned to one of two groups. As would be expected from a counterbalanced design, the groups differed in terms of the counterbalanced sequence in which they received the treatments. Half the participants (the group getting the sentence–image sequence) formed sentences for the first list, then formed images to recall the second list. The other half (the group getting the image–sentence sequence) formed images to recall the first list, then formed sentences to recall the second list.

Now that you have a basic understanding of the study's design, let's examine the study's results. To do so, look at both the table of means for that study (Table 13.8) and the analysis of variance summary table (Table 13.9).

TABLE 13.8
Table of Means for a Counterbalanced Memory Experiment

	MEMORY STRATEGY		
GROUP'S SEQUENCE	IMAGES	SENTENCES	IMAGES–SENTENCES DIFFERENCE
Group 1 (images first, sentences second)	8	6	+2
Group 2 (sentences first, images second)	6	8	−2
	14/2 = 7	14/2 = 7	Strategy Main Effect = 0

Counterbalancing Main Effect = 0

On the average, participants in both groups remembered a total of 14 words (8 in one condition, 6 in another)

Strategy Effect = 0

Average recalled in image condition was 7 ([8 + 6]/2).

Average recalled in sentence condition was 7([6 + 8]/2).

Order Effect = +2

Participants remember the first list best.

They averaged 8 words on the first list, 6 on the second.

The order (first vs. second) effect is revealed by an *interaction* involving counterbalancing *group* and rehearsal *strategy*.

That is, Group 1 did better in the image condition (8 to 6), but Group 2 did better in the sentence condition (8 to 6).

TABLE **13.9**
ANOVA Summary Table for a Counterbalanced Design

ANALYSIS OF VARIANCE TABLE					
SOURCE	*SS*	*df*	*MS*	*F*	*P*
Group Sequence (counterbalancing)	0	1	0	*0*	*n.s.**
Error Term for Between-Subjects	44	22	2		
Factor Memory Strategy	0	1	0	0	*n.s.*
Interaction Between Memory Strategy and Group Sequence	10	1	10	10	$p < .01$
(effect of order—first vs. second list) Within-Subjects Error Term	23	23	1.0		

**n.s.* is an abbreviation for not statistically significant.
Note: "*p*" values in an ANOVA summary table indicate the probability that the researchers could get differences between their conditions that were this big even if the variables were not related. That is, the *p* values tell you the probability that the difference between the groups could occur due to chance alone. Thus, the smaller the *p* value, the less likely the results are due only to chance—and the more likely that the variables really are related.

By looking at Table 13.9, we see that the main effect for the between-subjects factor of counterbalanced sequence is not significant. As Table 13.8 shows, members of both groups recalled, on the average, 14 words in the course of the experiment. Participants getting the treatment sequence A–B did not, on the average, recall more words than participants getting the sequence B–A.

Next, we see that the within-subjects factor of the memory strategy factor was also not significant. Because participants recalled the same number of words in the imagery condition (7) as they did in the sentence condition (7), we have no evidence that one strategy is superior to the other. Thus, there is no treatment effect.

Finally, we have a significant interaction of memory strategy and group sequence. By looking at Table 13.8, we see that this interaction is caused by the fact that Group 1 (which gets images first) recalled more words in the imagery condition whereas Group 2 (which gets sentences first) recalled more words in the sentences condition. In other words, participants do better on the first list than on the second.

What does this order (trials) effect mean? If the researchers were not careful in their selection of lists, the order effect could merely reflect the first list being made up of words that were easier to recall than the second list. The researchers, however, should not have made that mistake.[7] Therefore, if the experiment were properly conducted, the order effect must reflect either

[7] There are at least three ways to avoid this mistake: (a) extensively pretest the lists to make sure that both are equally memorable, (b) consult the literature to find lists that are equally memorable, and (c) counterbalance lists so that, across participants, each list occurred equally often under each instructional condition. The third approach is probably the best.

the effects of practice, fatigue, treatment carryover, or sensitization. In this case, it probably reflects the fact that the practice participants get on the first list hurts their memory for the second list. Psychologists do not consider this negative practice effect a nuisance. On the contrary, this negative practice effect is one of the most important and most widely investigated facts of memory—proactive interference.

Now that you understand the three effects (two main effects and the treatment × counterbalancing interaction) that you can find with a 2 × 2 counterbalanced design, let's look at an experiment where the researcher is interested in all three effects. Suppose that Mary Jones, a politician, produces two commercials: an emotional commercial and a rational commercial. She hires a psychologist to find out which commercial is most effective so she'll know which one to give more airtime. The researcher uses a counterbalanced design to address the question (see Table 13.10).

By looking at the treatment main effect, the researcher is able to answer the original question, "Which ad is more effective?" By looking at the counterbalancing sequence main effect, the researcher is able to find out whether one sequence of showing the ads is better than another, thus enabling him to answer the question, "Should we show the emotional ad first and then the rational ad or should we show the ads in the opposite sequence?" Finally, by looking at the ad × counterbalancing interaction, the researcher is able to determine if there is an order (trials) effect, leading him to be able to answer the question, "Are participants more favorable toward the candidate after they've seen the

TABLE **13.10**

Effects Revealed by a 2 × 2 Counterbalanced Design

GROUP 1	
First Ad	**Second Ad**
Emotional Ad	Rational Ad

GROUP 2	
First Ad	**Second Ad**
Rational Ad	Emotional Ad

Questions Addressed by the Design:

1. Is the rational ad more effective than the emotional ad? (Main effect of the within-subjects factor of type of ad)
2. Is it better to show the emotional ad and then the rational ad or the rational ad and then the emotional ad? (Main effect of the between-subjects factor of counterbalancing sequence)
3. Are attitudes more favorable to the candidate after seeing the second ad than after seeing the first? (Ad by counterbalancing interaction)

second ad?" Obviously, he would expect that voters would rate the candidate higher after seeing the second ad than they did after seeing the first ad.

Let's suppose that all three effects were statistically significant and the means were as follows:

	TYPE OF AD	
	Emotional Ad	**Rational Ad**
Group 1: (Emotional–Rational sequence)	4	6
Group 2: (Rational–Emotional sequence)	8	7

Note: Scores are rating of the candidate on a 1 (strongly disapprove of) to 9 (strongly approve of) scale.

As you can see from comparing the emotional ad column with the rational ad column, the treatment main effect is due to the rational ad, on the average, being more effective than the emotional ad. As you can see from comparing the Group 1 row with the Group 2 row, Group 2 likes the candidate more than Group 1. Thus, the between-groups counterbalancing sequence main effect suggests that it would be better to present the ads in the Rational–Emotional sequence (Group 2's sequence) than in the Emotional–Rational sequence (Group 1's sequence).

The table doesn't make it as easy to see the order effect. To see whether participants liked the candidate better after the second trial than after the first, this table makes you interpret the treatment × counterbalancing interaction. To help you find the order effect in this table, we have underlined the mean for the ad that each group saw first. Thus, we underlined 4 because Group 1 saw the emotional ad first, and we underlined 7 because Group 2 saw the rational ad first. By recognizing that $4 + 7$ is less than $8 + 6$, you could determine that scores were lower on the first trial than on the second. To better see the order effect, you should rearrange the table so that the columns represent "Order of Ads" rather than "Type of Ad." Your new table would look like this:

	ORDER OF ADS	
	First Ad	**Second Ad**
Group 1: (Emotional–Rational sequence)	4	6
Group 2: (Rational–Emotional sequence)	7	8

As you can see from this table, the order effect reveals that people like the candidate more after the second ad. The ads *do* build on each other.

It's possible, however, that the consultant may not have obtained an order effect. For example, suppose the consultant obtained the following pattern of results:

	TYPE OF AD	
	Emotional Ad	**Rational Ad**
Group 1: (Emotional–Rational sequence)	5	6
Group 2: (Rational–Emotional sequence)	5	6

In this case, Group 1 participants (who get the rational ad last) and Group 2 participants (who get the rational ad first) both rate the candidate one point higher after seeing the rational ad than they do after seeing the emotional ad. Thus, there is no treatment by counterbalancing interaction. Because there is no treatment × counterbalancing interaction, there is no order effect. However, an easier way to see that there was no order effect would be to create the following table.

	ORDER OF ADS	
	First Ad	**Second Ad**
Group 1: (Emotional–Rational sequence)	5	6
Group 2: (Rational–Emotional sequence)	6	5

With these data, the consultant would probably decide to just use the rational ad.

Instead of obtaining no order effect, the consultant could have obtained an order effect such that people always rated the candidate worse after the second ad. For example, suppose the consultant obtained the following results:

	ORDER OF ADS	
	First Ad	**Second Ad**
Group 1: (Emotional–Rational sequence)	5	4
Group 2: (Rational–Emotional sequence)	6	4

If the consultant obtained these results, he would take a long, hard look at the ads. It may be that both ads are making people dislike the candidate, or it may be that the combination of these two ads does not work. Seeing both ads may reduce liking for the candidate by making her seem inconsistent. For example, one ad may suggest that she supports increased military spending while the other may suggest that she opposes increased military spending.

Conclusions About Counterbalanced Within-Subjects Designs

As you can see from this last example, the counterbalanced design does more than balance out routine order effects. It also tells you about the impact of

both trials (order: first vs. second) and sequence (e.g., rational then emotional ad vs. emotional ad then rational ad). Therefore, you should use counterbalanced designs when

1. you want to make sure that routine order effects are balanced out
2. you are interested in sequence effects
3. you are interested in order (trials) effects

You will usually want to balance out order effects because you don't want order effects to destroy your study's internal validity. That is, you want a significant treatment main effect to be due to the treatment, rather than to order effects.

You will often be interested in **sequence effects** because real life is often a sequence of treatments (Greenwald, 1976). That is, most of us are not assigned to receive either praise or criticism; to see either ads for a candidate or against a candidate; to experience only success or failure, pleasure or pain, and so on. Instead, we usually receive both praise and criticism, see ads for and against a candidate, and experience both success and failure. Counterbalanced designs allow us to understand the effects of receiving different sequences of these "treatments." In counterbalanced designs, the main effect for the between-subjects factor of counterbalancing sequence can help you answer questions like the following:

- Would it be better to eat and then exercise—or to exercise and then eat?
- Would it be better to meditate and then study—or to study and then meditate?
- If you are going to compliment and criticize a friend, would you be better off to criticize, then praise—or to praise, then criticize?

Order (trials) effects, on the other hand, will probably interest you if you can control whether a particular event will be first or last in a series of events. Thus, you might be interested in using a counterbalanced design to find out whether it's best to be the first or the last person interviewed for a job. Or, if you want to do well in one particular course (research methods, of course), should you study the material for that course first or last? To find out about these order effects, you'd use a counterbalanced design and look at the treatment × counterbalancing interaction.

CHOOSING THE RIGHT DESIGN

If you want to compare two levels of an independent variable, you have several designs you can use: matched pairs, within-subjects designs, counterbalanced designs, and the simple between-subjects design. To help you choose among these designs, we will briefly summarize the ideal situation for using each design.

Choosing a Design When You Have One Independent Variable

The matched-groups design is ideal when

1. you can readily obtain participants' scores on the matching variable without arousing their suspicions about the purpose of the experiment

2. the matching variable correlates highly with the dependent measure
3. participants are scarce

The pure within-subjects design is ideal when

1. sensitization, practice, fatigue, or carryover effects are not problems
2. you want a powerful design
3. participants are scarce
4. you want to generalize your results to real-life situations, and in real life, individuals tend to be exposed to both levels of the treatment

The 2 × 2 counterbalanced design is ideal when

1. you want to balance out the effects of order
2. you are interested in order effects, sequence effects, or both
3. you have enough participants to meet the requirement of a counterbalanced design
4. you are not concerned that being exposed to both treatment levels will alert participants to the purpose of the experiment

The pure between-subjects design is ideal when

1. you think fatigue, practice, sensitization, or carryover effects could affect the results
2. you have access to a relatively large number of participants
3. you want to generalize your results to real-life situations, and in real life, individuals tend to receive either one treatment or the other, but not both

Choosing a Design When You Have More Than One Independent Variable

Thus far, we have discussed how to choose a design when you are studying the effects of a single variable (see Table 13.11). Often, however, you may want to investigate the effects of two or more variables.

In that case, you would appear to have three choices: a between-subjects factorial design, a within-subjects factorial design, and a counterbalanced design. However, counterbalancing becomes less attractive—especially for the beginning researcher—as the design becomes more complicated. Thus, as a general rule, beginning researchers who plan on manipulating two independent variables usually are choosing between a two-factor within-subjects design and a two-factor between-subjects design.

Using a Within-Subjects Factorial Design

You should use a pure within-subjects design when

1. you can handle the statistics (you will have to use within-subjects analysis of variance or multivariate analysis of variance)
2. sensitization, practice, fatigue, and carryover effects are not problems
3. you are concerned about power
4. in real-life situations, people are exposed to all your different combinations of treatments

TABLE **13.11**
Ideal Situations for Different Designs

SIMPLE EXPERIMENT	MATCHED GROUPS	WITHIN-SUBJECTS	COUNTERBALANCED DESIGN
Participants are plentiful.	Participants are very scarce.	Participants are very scarce.	Participants are somewhat scarce.
Order effects could be a problem.	Order effects could be a problem.	Order effects are not a problem.	Want to assess order effects or order effects can be balanced out.
Power isn't vital.	Power is vital.	Power is vital.	Power is vital.
In real life, people usually only get one or the other treatment, rarely get both.	In real life, people usually only get one or the other treatment, rarely get both.	In real life, people usually get both treatments, rarely get only one or the other.	In real life, people usually get both treatments, rarely get only one or the other.
Multiple exposure to dependent measure will tip participants off about hypothesis.	Exposure to matching variable will *not* tip participants off about hypothesis.	Multiple exposure to dependent measure will *not* tip participants off about hypothesis.	Multiple exposure to dependent measure will *not* tip participants off about hypothesis.
Exposure to different levels of the independent variable will tip participants off about hypothesis.	Exposure to different levels of the independent variable will tip participants off about hypothesis. Matching variable is easy to collect and correlates highly with the dependent measure.	Exposure to different levels of the independent variable will *not* tip participants off about hypothesis.	Exposure to different levels of the independent variable will *not* tip participants off about hypothesis.

Using a Between-Subjects Factorial Design

On the other hand, you should use a between-subjects design when

1. you are worried about the statistics of a complex within-subjects design
2. you are worried that order effects would destroy the internal validity of a within-subjects design
3. you are not worried about power
4. in real-life situations, people are exposed to either one combination of treatments or another

Using a Mixed Design

Sometimes, however, you will find it difficult to choose between a completely within-subjects design and a completely between-subjects design. For example, consider the following two cases.

> Case 1: You are studying the effects of brain lesions and practice on how well rats run mazes. On the one hand, you do not want to use a completely within-subjects design because you consider brain damage to occur "between subjects" in real life (because some individuals suffer brain damage and others do not). On the other hand, you do not want to use a completely between-subjects design because you

TABLE **13.12**
Ideal Situations for Making a Factor Between or Within

Should a Factor Be a Between-Subjects Factor or a Within-Subjects Factor?	
MAKE FACTOR BETWEEN SUBJECTS	MAKE FACTOR WITHIN SUBJECTS
Order effects pose problems.	Order effects are not a problem.
Lack of power is *not* a concern.	Lack of power is a serious concern.
You want to generalize the results to situations in which participants receive either one treatment or another.	You want to generalize the results to situations in which participants receive all levels of the treatment.

think that practice occurs "within subjects" in real life (because all individuals get practice and, over time, the amount of practice an individual gets increases).

Case 2: You are studying the effects of subliminal messages and electroconvulsive therapy on depression. You expect that if subliminal messages have any effect, it will be so small that only a within-subjects design could detect it. However, you feel that electroconvulsive shock should not be studied in a within-subjects design because of huge carryover effects (see Table 13.12).

Fortunately, in these cases, you are not forced to choose between a totally within-subjects factorial and a totally between-subjects factorial. As you know from our discussion of counterbalanced designs, you can do a study in which one factor is varied between subjects and the other is varied within subjects. Such designs, called **mixed designs**, are analyzed using a mixed analysis of variance. (To learn how to interpret the results of a mixed analysis of variance, see Box 13.4.)

In both Case 1 and Case 2, the mixed design allows us to have both internal validity and power. In Case 1, we could make lesions a between-subjects variable by randomly assigning half the participants to get lesions and half not. That way we do not have to worry about carryover effects from the brain lesions. We could make *practice* a within-subjects variable by having each participant run the maze three times. Consequently, we have the power to detect subtle differences due to practice (see Table 13.13 and Figure 13.1).

In Case 2, we could make ECS therapy a between-subjects variable by randomly assigning half the participants to get electroconvulsive (ECS) therapy and half not. That way, we do not have to worry about carryover effects from the ECS. Then, we would expose all participants to a variety of subliminal messages, some designed to boost mood and some to be neutral. By comparing the average overall depression scores from the ECS therapy group to that of the no-ECS group, we could assess the effect of ECS. By comparing participants' scores following the "positive" subliminal messages to their scores following "neutral" subliminal messages, we could detect even rather subtle effects of subliminal messages.

In a mixed design, you are able to test not only the main effects of two treatments but also the interaction of those treatments. In Case 1, the interesting statistical effects will probably involve the interaction rather than the two

BOX 13.4 Not Getting Mixed Up About Mixed Designs

If you use a mixed design, you will probably have a computer analyze your data for you. Often, both entering the data and interpreting the printout are straightforward. For example, suppose you had two groups (one received Treatment X, the other Treatment Y), had each participant go through three trials, and collected the following data.

PARTICIPANT	GROUP	TRIAL 1	TRIAL 2	TRIAL 3
Steve	X	1	3	7
Mary	X	2	4	6
Todd	X	3	6	7
Melissa	X	4	5	7
Tom	Y	4	5	7
Amy	Y	5	4	7
Rob	Y	4	5	6
Kara	Y	4	4	7

You might input the data as follows.

GROUP	TRIAL 1	TRIAL 2	TRIAL 3
1	1	3	7
1	2	4	6
1	3	6	7
1	4	5	7
2	4	5	7
2	5	4	7
2	4	5	6
2	4	4	7

Your printout might be relatively straightforward and resemble the following.

	T 1 MEAN	T 2 MEAN	T 2 MEAN
Group 1	2.5	4.5	6.75
Group 2	4.25	4.5	6.75
Total	3.375	4.5	6.75

BETWEEN Ss

Source	*SS*	*df*	*MS*	*F*	*p*
A	2.04	1	2.04	1.69	.24
Error term	7.25	6	1.21		

WITHIN Ss

B	47.25	2	23.63	47.26	<.001
A × B	4.08	2	2.04	4.08	.044
Error term	6.0	12	.5		

However, in some programs, entering your data and interpreting the printout can be more complicated. To make sure that the computer has done the analysis you expected, check your printout carefully.

If your printout contains only one error term, the computer is analyzing your data as if you have a completely between-subjects design. If you take the *MS* for any treatment or interaction and divide it by your one and only *MSE,* you will get the *F* for that effect.

If, on the other hand, every main effect and every interaction has its own error term, the computer is analyzing your data as if you have a completely within-subjects design. In that case, if you have three effects (two main effects and an interaction effect), you will have three error terms.

Even if the computer seems to be analyzing your study as a mixed design, check the computer printout to be sure that it has correctly identified which factors are within and which are between. Start by looking at the degrees of freedom for all your main effects. If your between-subjects factor(s) have more levels than your within-subjects factor(s), then the degrees of freedom for your between-subjects main effect should be larger than the degrees of freedom for your within-subjects main effect. In any event, make sure that the *df* for each of your variable's main effects is one fewer than the number of levels of that variable. For example, if you have 4 levels of the between variable and 2 levels of the within variable, be sure that the degrees of freedom for the between variable is 3 and that the degrees of freedom for the within variable is 1.

Next, focus on your between-subjects factor(s). All between-subjects main effects—and all

(Continued)

BOX 13.4 Continued

interactions that involve only between-subjects factors—should be tested against a single error term. To guarantee this, divide the *MS* for each between factors main effect and each exclusively between factors interaction by the *MS* for the between-subjects error term. In every case, you should get the same *F* that is reported in the printout.

To double-check that the computer correctly identified all the between-subjects variables, add up the degrees of freedom for all the between-subjects main effects, the *df* for the interactions that involved only between-subjects factors, and the *df* for the between-subjects error term. The total of these degrees of freedom should be one fewer than the number of participants.

Next, check the within factors. Each within-subjects main effect and each interaction that involves only within-subjects factors should be tested against a different error term.

Finally, look at interactions in which at least one variable is a between factor and at least one variable is a within factor. To find the appropriate error term for these interactions, attend only to the within-subjects factors: Ignore the between-subjects factors. If A is a within factor and B is a between factor and you see an A × B interaction, this interaction should be tested against the same error term that A is tested against. If A is a within factor and B and C are between factors, the error term for the A × B × C interaction should still be the same error term that was used for testing A. If it is not, there is a mix-up about which of your factors are within and which are between.

TABLE **13.13**
Analysis of Variance Summary Table for a Mixed Design

SOURCE OF VARIANCE	*dF*	*SS*	*MS*	*F*	*P*
Brain Lesion	1	51.0	*51.0*	*10.0*	.0068
Between-Subjects Error	14	72.4	5.1		
Trials	2	26.6	13.3	11.1	.0003
Lesions × Trials	2	13.7	6.8	5.7	.0083
Within-Subjects Error	28	33.6	1.2		

Note: The mean square error for the within-subjects term is much smaller than the between-subjects error term (1.2 to 5.1), giving the design tremendous power for detecting within-subjects effects. This table corresponds to the graph in Figure 13.1.

main effects. That is, we would not be terribly surprised to find a main effect for lesion, telling us that the brain-lesioned rats performed worse.[8] Nor would we be surprised to find a main effect for practice, telling us that participants improve with practice. However, we would be interested in knowing about the practice × lesion interaction. A significant practice × lesion interaction would tell us that one group of rats was benefiting from practice more than another. In this case, as you can see from Figure 13.1, the nonlesion

[8] The lesion main effect would be especially unsurprising if our control group didn't get any surgery. However, such empty control groups are rare. Typically, the control group would be a "sham lesion" control group that got brain surgery and was treated the same as the treatment group except that, instead of being injected with a chemical that would destroy (lesion) part of the brain, they would be injected with a harmless saline solution.

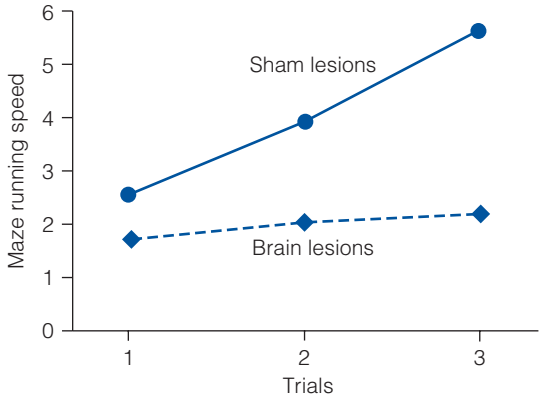

FIGURE **13.1** An Interaction in a Mixed Design

group benefits most from practice. In Case 2, although we would be interested in both the ECS and subliminal message main effects, we might be most interested in the interaction between ECS and subliminal messages: Such an interaction would tell us whether the ECS group was more influenced by the subliminal messages than the no-ECS group.

In many mixed designs, both a main effect and the interaction will be of interest. For example, Hebl and Mannix (2003) found a between-subjects main effect indicating that participants who saw a picture of a male job applicant sitting next to an overweight woman rated the job applicant more harshly than participants who saw a picture of the same man sitting next to an average-weight woman. This between-subjects main effect was of interest. The interaction between this main effect and the within-subjects variable of rating dimension (willingness to hire applicant, applicant's professional qualities, applicant's interpersonal skills) was also of interest because Hebl and Mannix wanted to see whether being seen with an overweight woman influenced hiring judgments more than it affected judgments about the applicant's interpersonal skills.

Note the problems Hebl and Mannix would have had in interpreting their results if they had used either a completely within-subjects or a completely between-subjects design. If they had used a completely within-subjects design, each participant would rate the applicant both (1) after seeing the applicant in the presence of an overweight woman and (2) after seeing the applicant in the presence of an average-weight woman. Participants would have found the study strange and would probably have figured out the hypothesis, thereby making the weight of woman main effect hard to interpret.

If Hebl and Mannix had used a completely between-subjects design, one group of participants would make hiring judgments, another group would make interpersonal skills judgments, and yet another group would make judgments about the applicant's professional qualities. Because each participant would be providing one set of ratings rather than the three sets that Hebl and Mannix's participants did, each participant in a between-subjects design would be providing only 1/3 as much data as the participants in Hebl and Mannix's actual study. Because participants would be providing less

data, the study would have been less powerful than Hebl and Mannix's actual study. Thus, if Hebl and Mannix had used a completely between-subjects design and failed to find an effect for the interaction, a scientist reading their work would wonder whether they would have succeeded in finding an interaction had they used a more powerful design.

As you can see from Hebl and Mannix's study and from our two hypothetical cases (Case 1 and Case 2), the mixed design has two major strengths. First, it allows you to examine the effects of two independent variables and their interaction. Second, instead of trading off the needs of one variable for the needs of another, you are able to give both variables the design they need. Because of its versatility, the mixed design is one of the most popular experimental designs.

CONCLUDING REMARKS

This chapter has expanded your ability to read about and conduct research. When reading reports of either within-subjects or mixed designs, you now know to ask

1. whether the multiple measures and manipulations may have led participants to figure out the hypothesis
2. what steps (e.g., counterbalancing) were taken to reduce order effects (practice, fatigue, carryover, and sensitization)—and whether those steps were sufficient to ensure the study's internal validity
3. whether a between-subjects design might have been more internally valid

When planning, conducting, or analyzing research, you now can

1. do experiments to determine the effect of a treatment and have a reasonable chance of finding the treatment effect even if the effect is small and you have access to only a few participants
2. replicate between-groups experiments that failed to find an effect with a more powerful design that is more likely to find an effect
3. use counterbalancing to control for order effects
4. take steps to minimize practice, fatigue, carryover, and sensitization, thereby minimizing order effects
5. do research assessing the effects of order (trials) and the effect of interactions involving trials (e.g., does the effect of one treatment get stronger when it is repeatedly presented whereas the effect of another treatment weakens with repeated exposures?)
6. do research to determine the effect of different treatment sequences (e.g., is it more effective to have cognitive therapy followed by antidepressants or to have antidepressants followed by cognitive therapy?)
7. determine whether you should use a pure between-subjects experiment, a matched-pairs experiment, a within-subjects design, or a mixed design
8. interpret computer printouts of analysis of variance (ANOVA) analyses of within-subjects as well as mixed designs

SUMMARY

1. The matched-pairs design uses matching to reduce the effects of random differences between participants and uses random assignment and statistics to account for the remaining effects of random error. Because of random assignment, the matched-pairs design has internal validity. Because of matching, the matched-pairs design has power.
2. Because the matched-pairs design gives you power without limiting the kind of participant you can use, you may be able to generalize your results to a broader population than if you had used a simple experiment.
3. The matched-pairs design's weaknesses stem from matching: Matching may sensitize participants to your hypothesis and participants may drop out of the study between the time of the matching and the time the experiment is performed.
4. Within-subjects designs are also known as repeated-measures designs.
5. The two-condition within-subjects design gives you two scores per participant.
6. The within-subjects design increases power by eliminating random error due to individual differences and by increasing the number of observations that you obtain from each participant.
7. Both the matched-pairs design and the two-condition pure within-subjects design can be analyzed by the dependent groups *t* test. Complex within-subjects designs require more complex analyses. Specifically, they should be analyzed by within-subjects analysis of variance (ANOVA) or by multivariate analysis of variance (MANOVA).
8. Because of practice, fatigue, carryover, and sensitization effects, the participant may respond one way if receiving a treatment first and a different way if receiving the treatment last. These order effects may make it difficult to assess a treatment's real effect.
9. To reduce the effects of order, you should randomly determine the sequence in which each participant will get the treatments or use a counterbalanced design.
10. In the counterbalanced design, participants are randomly assigned to systematically varying sequences of conditions to ensure that routine order effects are balanced out.
11. Order effects (often called trials effects) are different from sequence effects. *Order effects* refer to whether participants respond differently on one trial (e.g., the first) than on some other trial (e.g., the last). Order is a within-subjects factor in a counterbalanced design.
12. Order effects can be detected by looking at the treatment × counterbalancing sequence interaction.
13. *Sequence effects* refer to whether participants respond differently to getting a series of treatments in one sequence than getting the treatments in a different sequence. For example, the group of participants who get the treatments arranged in the sequence Treatment A, then Treatment B may have higher overall average scores than the group of participants who get the treatments arranged in the sequence Treatment B, then Treatment A. Sequence is a between-subjects factor.
14. A counterbalanced design allows you to assess the effect of (a) the treatment, (b) receiving different counterbalanced sequences of treatments, and (c) order (whether participants respond differently on the first trial than on the last).
15. Because you must include the between-subjects factor of counterbalancing in your analyses, counterbalanced designs require more participants than pure within-subjects designs.
16. If you want to compare two levels of an independent variable, you can use a matched-pairs design, a within-subjects design, a counterbalanced design, or a simple between-subjects design.
17. Mixed designs have both a within- and a between-subjects factor. Counterbalanced designs are one form of a mixed design.
18. Mixed designs should be analyzed with a mixed analysis of variance or a multivariate analysis of variance.

KEY TERMS

mixed designs (*p. 465*)
matched-pairs design (*p. 466*)
power (*p. 467*)
dependent groups *t* test (*p. 471*)
within-subjects designs (*repeated-measures designs*) (*p. 474*)
order (*p. 475*)
order (trial) effects (*p. 476*)
practice effects (*p. 477*)
fatigue effects (*p. 477*)
carryover (treatment carryover) effects (*p. 477*)
sensitization (*p. 477*)
randomized within-subjects design (*p. 481*)
counterbalanced within-subjects design (*p. 483*)
sequence effects (*p. 493*)

EXERCISES

1. What feature of the matched-pairs design makes it
   a. an internally valid design?
   b. a powerful design?
2. A researcher uses a simple between-subjects experiment involving 10 participants to examine the effects of memory strategy (repetition vs. imagery) on memory.
   a. Do you think the researcher will find a significant effect? Why or why not?
   b. What design would you recommend?
   c. If the researcher had used a matched-pairs study involving 10 participants, would the study have more power? Why? How many degrees of freedom would the researcher have? What type of matching task would you suggest? Why?
3. An investigator wants to find out whether hearing jokes will allow a person to persevere longer on a frustrating task. The researcher matches participants based on their reaction to a frustrating task. Of the 30 original participants, 5 quit the study after going through the "frustration pretest." Beyond the ethical problems, what problems are there in using a matched-pairs design in this situation?
4. What problems would there be in using a within-subjects design to study the "humor-perseverance" study (discussed in question 3)? Would a counterbalanced design solve these problems? Why or why not?
5. Why are within-subjects designs more powerful than matched-pairs designs?
6. Two researchers hypothesize that spatial problems will be solved more quickly when the problems are presented to participants' left visual fields than when stimuli are presented to participants' right visual fields. (They reason that messages seen in the left visual field go directly to the right brain, which is often assumed to be better at processing spatial information.) Conversely, they believe verbal tasks will be performed more quickly when stimuli are presented to participants' right visual fields than when the tasks are presented to participants' left visual fields. What design would you recommend? Why?
7. A student hypothesizes that alcohol level will affect sense of humor. Specifically, the student has two hypotheses. First, the more people drink, the more they will laugh at slapstick humor. Second, the more people drink, the less they will laugh at other forms of humor. What design would you recommend the student use? Why?
8. You want to determine whether caffeine, a snack, or a brief walk has a more beneficial effect on mood. What design would you use? Why?
9. Using a driving simulator and a within-subjects design, you want to compare the differences between driving unimpaired, driving while talking on a cell phone, and driving while legally intoxicated.
   a. Which order effects do you have to worry about? Why?

   b. To what degree would counterbalancing solve the problems caused by order effects?
   c. How would you try to prevent order effects from harming the validity of your study?

10. A researcher wants to know whether music lessons increase scores on IQ subtests and whether music lessons have more of an effect on some subtests (e.g., more of an effect on math than on vocabulary) than others.
   a. Would you make music lessons a between- or within-subjects factor? Why?
   b. Would you make subtests a between- or within-subjects factor? Why?
   c. If the researcher did an analysis of variance (ANOVA) on the data, the researcher would obtain three effects. Name those three effects.
   d. What effect would the researcher look for to determine whether music lessons increase scores on IQ subtests?
   e. What effect would the researcher look for to determine whether music lessons have more of an effect on math subtests than on vocabulary subtests?

WEB RESOURCES

Go to the Chapter 13 section of the book's student website and

1. Look over the concept map of the key terms.
2. Test your self on the key terms.
3. Take the Chapter 13 Practice Quiz.
4. Download the Chapter 13 tutorial to practice
   a. distinguishing between order and sequence effects
   b. interpreting printouts from within-subjects designs
   c. choosing among designs
5. Do an analysis on data from a within-subjects design using a statistical calculator by going to the "Statistical Calculator" link.