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

CHAPTER 11

Experimental Design: Adding to the Basic Building Block

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

In Chapter 10 we learned many concepts and principles about **experimental design** that are basic to planning *any* experiment, not merely the basic two-group experiment. When we come to one of those topics in this chapter, we will briefly review it and refer you back to Chapter 10 for the original discussion.

In this chapter we will add to our basic building-block design. Consider our previous analogy: As a child you quickly mastered the beginner's set of Legos or Tinkertoys. You learned to build everything there was to build with that small set and then wanted to go beyond those simple objects to build larger, more exciting creations. To satisfy this desire, you got a larger set of building materials that you could combine with the starter set in order to build more complicated objects. Despite the fact you were using a larger set of materials, the basic principles you learned with your starter set still applied.

Experimental design works in much the same way. Researchers typically want to move beyond two-group designs so that they can ask more complicated, more interesting questions. Fortunately, they don't have to start from scratch—that's why we referred to the two-group design as the basic building-block design in the previous chapter. Every experimental design is based on the two-group design. Although the questions you ask may become more complicated or sophisticated, your experimental design principles will remain constant. In the same way, when they face a more difficult case, detectives continue to use the basic investigative procedures they have learned.

It is still appropriate to think of your experimental design as the blueprint for your experiment. We hope the following analogy convinces you of the need for having an experimental design. Although you *might* be able to get by without a blueprint if you're building a doghouse, it is unlikely you would want to build your own house without a blueprint. Think of building a small house as being equivalent to using the two-group design from Chapter 10. If you need a blueprint to build a house, imagine how much more you would need a blueprint to build an apartment building or a skyscraper. We will work toward the skyscrapers of experimental design in Chapter 12.

The Multiple-Group Design

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

Here, we will consider an extension of the two-group design. Turn back to Figure 10-2 (page 207) for just a moment. What would be the next logical step to add to this design so that we could ask (and answer) slightly more complex questions?

How Many IVs? The first question that we ask when considering any experimental design is always the same: "How many **independent variables (IVs)** will I use in my experiment?" (see Figure 11-1). In this chapter we will continue to consider only experiments that use one IV. We should remember that although one-IV

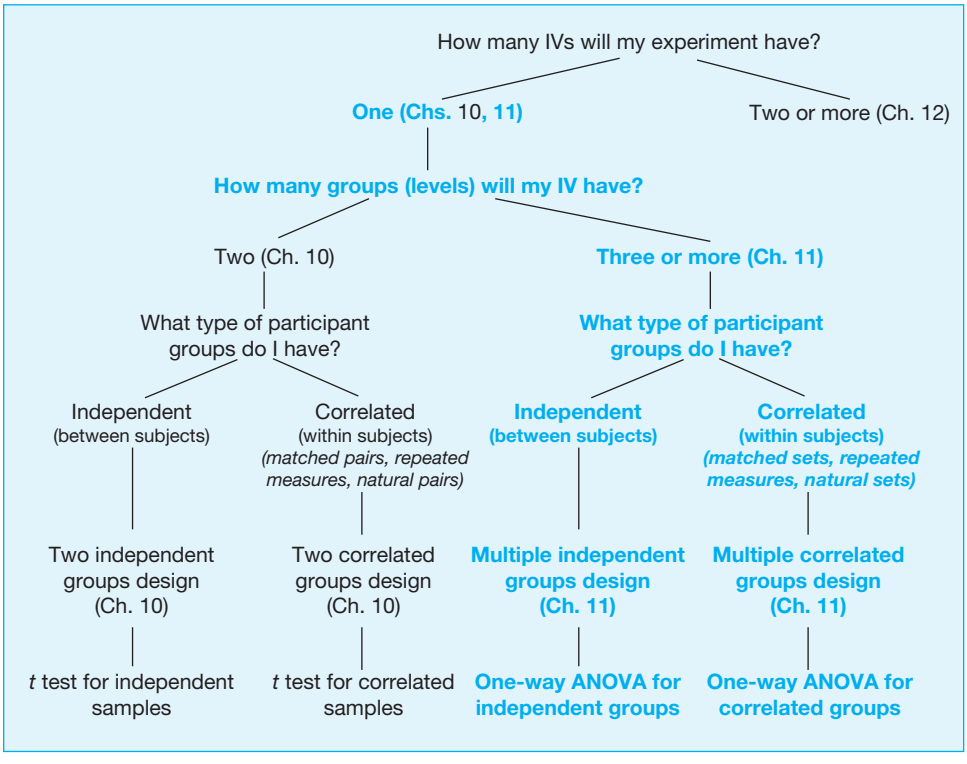


FIGURE 11-1 Experimental Design Questions.

experiments may be simpler than experiments that use multiple IVs (Chapter 12), they are not inferior in any way. Many students who are designing their first research study decide to throw in everything except the kitchen sink as an IV. A well-designed experiment with one IV is vastly preferable to a sloppy experiment with many variables thrown together. Remember the **principle of parsimony** from Chapter 10: If a one-IV experiment can answer your questions, use it and don't complicate things needlessly.

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

How Many Groups? As soon as we have decided to conduct a one-IV experiment, our second question (see Figure 11-1) revolves around how many groups we will use to test the IV. This question marks the difference between the multiple-group design and the two-group design. As their names imply, a two-group design compares two **levels** of an IV, whereas a multiple-group design compares three or more levels of a single IV. Thus, we could compare three, four, five, or even more differing levels or amounts of an IV. This experimental situation is similar to that experienced by a detective faced with multiple suspects. Instead of merely investigating two people, the detective must conduct simultaneous investigations of more than two individuals.

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

In Chapter 10 we learned that the most common type of two-group design uses experimental and control groups. How will the multiple-group design differ from that common two-group design?

Actually, there are two answers to that question—if you got either, pat yourself on the back. First, a multiple-group design *can* have a control group. Rather than having a single experimental group and a control group, a multiple-group design with a control group would also have two or more experimental groups. This combination allows us to condense several two-group experiments into one experiment. Instead of conducting a two-group experiment to determine whether your IV has an effect and a second two-group experiment to determine the optimum amount of your IV, you could conduct a multiple-group experiment with a control group and the number of **treatment groups** you would like to assess. Second, a multiple-group design does not have to have a control group. If you already know that your IV has an effect, you can simply compare as many treatment groups as you would like in your multiple-group design.

Treatment groups Groups of participants that receive the IV.

Let's look at a research example using the multiple-group design. Colleen Sullivan and her faculty advisor, Camille Buckner (2005), of Frostburg State University in Frostburg, Maryland, wanted to determine whether the type of role model affected students' intentions to purchase products. They used parental, peer, celebrity, and no role model conditions. Why does this experiment fit the multiple-group design? First, it has one IV: the type of role model. Second, the IV has more than two levels: It has four, based on the four different role model conditions. Thus, as you can see in Figure 11-1, with one IV and four levels, this experiment requires a multiple-group design. We can draw the block diagram depicted in Figure 11-2 to

	INDEPENDENT VARIABLE (TYPE OF ROLE MODEL)		
EXPERIMENTAL GROUP 1	EXPERIMENTAL GROUP 2	EXPERIMENTAL GROUP 3	EXPERIMENTAL GROUP 4
Parent	Peer	Celebrity	No role model

FIGURE 11-2 The Multiple-Group Design Used by Sullivan and Buckner (2005).

Source: From "The Influence of Perceived Role Models on College Students' Purchasing Intention and Product-Related Behavior," by C. J. Sullivan and C. E. Buckner, 2005, *Psi Chi Journal of Undergraduate Research, 10,* pp. 66–71.

portray this experimental design. By comparing Figures 10-2 and 11-2, you can easily see how the multiple-group design is an extension of the two-group design.

Three of the groups shown in Figure 11-2 are experimental groups. Does this experiment use a control group? Yes, the *no role model* condition served as a control group. In this experiment, Sullivan and Buckner were interested in the various differences among the four groups based on their differing types of role models. They found that students in the parental role model condition were higher in their intention to buy products than students in the celebrity and no role model conditions; students in the peer role model condition scored between the parental and other two conditions but were not significantly different from any group. In statistical terminology, then, Sullivan and Buckner found results that supported the experimental hypothesis (i.e., that there was a difference between the performance of the groups as a function of role model condition). Of course, support for the experimental hypothesis is not the same as *proving* the experimental hypothesis. Did Sullivan and Buckner *prove* that the type of role model affects students' buying intentionality? No, they merely demonstrated that there was a difference that was unlikely to have occurred by chance for the role model conditions they used and for their groups of participants. What about using different role models? What about using different research participants—children or older adults, for example? Recall that in Chapter 8 we talked about how to generalize our results beyond the specific participants in our experiment.

Suppose you wished to test more than three conditions. Could you use a multiple-group design in such a case? Why or why not? If so, what would happen to the block design in Figure 11-2?

Yes, you could use the multiple-group design if you had four or five role models to assess. In fact, it could be used if there were 10 or 20 role models. The only requirement for using the multiple-group design is an experiment with one IV and more than two groups (see Figure 11-1). Practically speaking, it is rare that multiple-group designs are used with more than four or five groups. If we did use such a design with more than three groups, we would merely extend our block diagram, as shown in Figure 11-3.

Assigning Participants to Groups After we decide to conduct a multiple-group experiment, we must decide about the assignment of research participants to groups (see

INDEPENDENT VARIABLE (TYPE OF ROLE MODEL)

Expl. Group 1	Expl. Group 2	Expl. Group 3	Expl. Group 4	Expl. Group 5
Role model 1	Role model 2	Role model 3	Role model 4	Role model 5

FIGURE 11-3 Hypothetical Multiple-Group Design With Five Groups.

Figure 11-1). Just as in Chapter 10, we choose between using **independent groups** or **correlated groups**.

Independent Samples (Random Assignment to Groups) Remember that with **random assignment** each participant has an equal chance of being assigned to any group. In their experiment on the effects of role models on buying intention, Sullivan and Buckner (2005) used random assignment when they assigned students to groups: All 69 students had a 1 in 4 chance of being in the parental, peer, celebrity, or no role model group. When we use large groups of participants, random assignment should create groups that are equal on potential extraneous variables such as personality, age, and sex. Recall from Chapter 9 that random assignment allows us to control extraneous variables about which we are unaware. Thus, random assignment serves as an important **control procedure**. For example, we would not want to use role models with only male or female college students. We want to spread the different levels of the IV across all types of participants in order to avoid a **confounded experiment**. Suppose we put all women in the celebrity role model condition and all men in the parental role model condition. When we tabulated our results, we would not be able to draw a clear conclusion about the effects of type of role model because role model was confounded with participant sex. In other words, we couldn't be sure whether a significant difference between groups was caused by the role model difference between the groups *or* the sex difference between the groups.

Random assignment results in participants who have no relation to participants in other groups; in other words, the groups are independent of each other. We are interested in comparing differences between the various independent groups. As shown in Figure 11-1, when we use random assignment in this design, we end up with a multiple-independent-groups design.

Correlated Samples (Nonrandom Assignment to Groups) In the multiple-group design, we have the same concern about random assignment that we did with the two-group design: What if the random assignment does not work and we begin our experiment with unequal groups? We know that random assignment *should* create equal groups but also that it is most likely to work in the long run—that is, when we have many participants. If we have few participants or if we expect only small differences owing to our IV, we may want more control than random assignment affords us. In such situations, we often resort to using nonrandom methods of assigning participants to groups and thus end up with

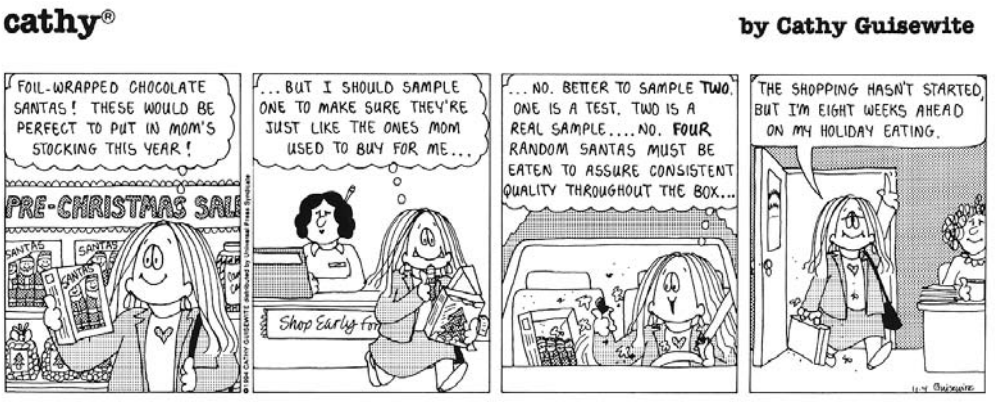
Independent groups Groups of participants formed by random assignment.

Correlated groups Groups of participants formed by matching, natural pairs, or repeated measures.

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

Control procedure One of several steps experimenters take to ensure that potential extraneous variables are controlled, including random assignment, matching, and so on.

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



Is Cathy using random assignment or random selection in this cartoon? You may want to look back at Chapters 6 and 7.

Because she is tasting chocolate Santas in a nonsystematic way (rather than assigning Santas to groups), Cathy's gluttony illustrates random selection.

correlated groups. Let's examine our three ways of creating correlated groups and see how they differ from the strategies discussed in Chapter 10.

1. ***Matched Sets.*** Matched *pairs* are not appropriate for the multiple-group design because we have at least three groups; therefore, we must use matched *sets*. The principle for forming matched sets is the same as that for forming matched pairs. Before our experiment we measure our participants on some variable that will affect their performance on the DV. Then we create sets of participants who are essentially the same on this measured variable, often known as the **matching variable**. The size of the sets will, of course, depend on how many levels our IV has. If our IV has five levels, for example, then each set would have five participants equated on the matching variable. After we create our matched sets, we then randomly assign the participants within each set to the different groups (treatments).

Matching variable A potential extraneous variable on which we measure our research participants and from which we form sets of participants who are equal on the variable.

Returning to Sullivan and Buckner's (2005) experiment, suppose we believed that participant sex would be an extraneous variable because we were using an attractive male for the celebrity role model condition. To ensure equality of their groups on sex, Sullivan and Buckner could have used sets of four participants who were matched on sex, with each participant then randomly assigned to one of the four groups. In this manner, they would have assured that the distribution of participant sex was uniform across all three groups. If all four of their groups had the same sex composition (regardless of whether it was 50-50), then participant sex could not be an extraneous variable.

One final caution is in order. We should remember that the potential extraneous variable must actually affect performance on the DV or else we have hurt our chances of finding a significant difference.

2. ***Repeated Measures.*** Other than the fact that participants perform in three or more conditions rather than only two, repeated measures in the multiple-group design are identical to repeated measures in the two-group design. When you use repeated measures in the multiple-group design, each participant must take part in *all* the various

treatment conditions. Thus, for Sullivan and Buckner to have used repeated measures, each student would have to have participated in the parental, peer, celebrity, and no role model conditions.

Can you see any possible flaws in conducting Sullivan and Buckner's experiment using a repeated-measures design? (Hint: Consider some of the practical issues from Chapter 10 about using repeated measures.)

Several problems could occur if we attempted the role modeling experiment with a repeated-measures design. Would you have one student complete the same survey four different times? This approach would not seem to make any sense. The students might become suspicious about the differing role models; they would probably be able to figure out what the IV was. Or they might get bored from completing the survey four times and not take it seriously on the third or fourth administration. What about using four different surveys? Again, this approach could be problematic. Could we logically assume that students gave different responses because of the role model? Not necessarily, because different surveys would have to differ in some way from one other. Students might give different types of answers to different types of surveys. Thus, we might be measuring the response to different surveys rather than to different types of role models! It seems that this experiment simply would not work well as a repeated-measures design. Remember that we mentioned this possible problem in Chapter 10—not all experiments can be conducted using repeated measures.

3. ***Natural Sets.*** Using natural sets is analogous to using natural pairs except that our sets must include more than two research participants. Using multiple groups takes away our ability to use some interesting natural pairs such as twins or husbands and wives, but other possibilities for using natural sets do exist. For example, many animal researchers use littermates as natural sets, assuming that their shared heredity makes them more similar than randomly selected animals. In a similar fashion, if your research participants were siblings from families with three (or more) children, natural sets would be a possibility. Most natural pairs or sets involve biological relationships.

We create multiple-correlated-groups designs when we use matched sets, repeated measures, or natural sets. The critical distinction is that the participants in these types of groups are related to each other in some way—we are comparing differences *within* groups (or within subjects, to use the old terminology). On the other hand, in independent-groups designs, the participants have no common relationship. Thus, we compare differences *between* differing groups of subjects.

Kimberly Walker and her faculty advisors James Arruda and Keegan Greenier (1999) of Mercer University in Macon, Georgia, conducted an experiment that used a multiple-correlated-groups design. They were interested in measuring how accurately people respond on the Visual Analogue Scale (VAS). The VAS is a self-report scale used for people to indicate internal states such as mood, hunger, or pain. Participants respond on the VAS by making a mark somewhere along a line, typically 100 millimeters (mm) in length, which has marks at the ends designating extreme low and extreme high values of the internal state. Walker et al.

asked participants to make marks at distances of 10, 20, 30, 40, 50, 60, 70, 80, and 90 mm along the line. They calculated participants' errors by subtracting the difference between the actual distance and the marked distance. By measuring the errors over nine distances, Walker et al. could compare data to determine whether errors were more likely to occur at different points along the line.

Why does the Walker, Arruda, and Greenier experiment illustrate a correlated-groups design? Which particular correlated-groups technique did they use? Why do you think they used a correlated-groups design?

In this experiment the IV was distance, measured in mm. Thus, there were nine levels of the IV. They measured each participant on each different distance; therefore, the correlated groups were a result of using repeated measures. Walker et al. probably used a repeated-measures design because it would take an enormous number of participants if they had each participant serve in only one distance group (nine groups of people needed). Also, it would have been an inefficient use of participants to use nine different groups. Finally, by having people participate in all levels of the IV, Walker et al. did not have to worry about the assumption that the groups were equivalent before the experiment began. By using each person as his or her own control across the nine measurements, the question of group equality was answered. The use of repeated measures helped control many subject variables that *might* have affected participants' performance on the VAS—factors such as motivation, spatial ability, and sex.

Walker et al. (1999) found that the participants were more accurate in making marks at the beginning and the end of the VAS line than in the middle of the line. They explained their results as being due to a perceptual phenomenon: the law of visual angle. Furthermore, they recommended adding depth cues to the VAS to increase accuracy on the task.

Comparing the Multiple-Group and Two-Group Designs

As in Chapter 10, we have to make a decision about how to design our potential experiment. Just as in Chapter 10, there are two multiple-group designs from which to choose. Researchers who want to design an experiment with one IV, however, must also choose between multiple-group designs and two-group designs. In the following sections, we will examine the various advantages and disadvantages of these experimental designs. As we warned you in Chapter 10, read carefully—you may be facing this choice yourself in the future!

The multiple-group design is quite similar to the two-group design. As a matter of fact, all you have to do to change your two-group design into a multiple-group design is add another level (or more) to your IV. Given this high degree of similarity, how would we compare these two designs?

In choosing a design for your experiment, your paramount concern is your experimental question. Does your question require only two groups to find an answer, or does it necessitate three or more groups? This question almost seems like a "no-brainer," but it cannot be taken for granted. Following the principle of parsimony from Chapter 10, we want to select the simplest possible design that will answer our question.

In Chapter 10 we provided you with an ideal situation for a two-group design—an experiment in which we merely wished to determine whether our IV has an effect. Often such an experiment is not necessary because that information already exists in the literature. You should never conduct an experiment to determine whether a particular IV has an effect without first conducting a thorough literature search (see Chapter 2). If you find no answer in a library search, then you should consider conducting a two-group (presence–absence) study. If, however, you find the answer to that basic question and wish to go farther, a multiple-group design might be appropriate.

After these considerations, what do you do when you face a situation in which either a two-group design or a multiple-group design is appropriate? Although this answer may sound odd, you should think about your (future) results. What will they look like? What will they mean? Most critically, what will the addition of any group(s) beyond the basic two tell you? If the information that you expect to find by adding a group or groups is important and meaningful, then by all means add the groups. If, however, you're not really certain what you might learn by adding to the two groups, then you may be merely complicating your experiment needlessly.

Think back to the two student examples cited in this chapter. Did the researchers learn important information by adding an extra group to their two groups? Sullivan and Buckner (2005) found that intent to buy a product differed with four different types of role models (parent, peer, celebrity, and no model). Some people might wish to examine more levels in this experiment rather than fewer. In fact, one of your first thoughts when you read about Sullivan and Buckner's experiment earlier may have been, "I wonder how students would respond to _____ as a role model?" (Fill in the blank with a person or entity whom you believe would be particularly persuasive, e.g., an athlete.) You may believe that this experiment was not a fair test of the question about buying based on role model, especially if you disagree with their role model choices. It appears that Sullivan and Buckner made a wise decision in using a multiple-group design rather than a two-group design. In fact, it might have been more informative had they used an even larger multiple-group design.

Walker et al. (1999) measured people's responses on the VAS to nine different distances. They clearly benefited by using a multiple-group design. If they had simply measured the responses on short or long distances, they would have found little error in the responses.

Suppose Walker et al. (1999) wanted to use a smaller or larger multiple-group design—say measuring people's responses over 5 or 15 distances rather than 9. Would it be possible to use such a small or large multiple-group design?

Could you use a multiple-group design with 5 or 15 groups or measurements? Of course you could. The only limitation would be a practical consideration: Could you secure enough participants to take part (for matched sets or natural sets), or can the participants cope with being measured so many times (for repeated measures)? Because the Walker et al. experiment used repeated measures, our concern there would be for the experimental participants. In this case the participants would be measured either fewer or more times. This experiment, then, would be no problem to run over 5 measurement distances; participants would have fewer responses

to make. The question is whether 5 distances would give the experimenters all the information they needed. On the other hand, using 15 measurements might make the task extremely difficult for the participants to discriminate, and it would require more of the participants' time. Using nine levels of an IV for repeated measures is quite unusual, so it is much more likely that other experimenters would use smaller, rather than larger, experimental designs.

In summary, the multiple-group and two-group designs are quite similar; however, there are important differences between them that you should consider when choosing an experimental design for your research project.

Comparing Multiple-Group Designs

As you might guess, our comparison of the multiple-independent-groups design to the multiple-correlated-groups design is going to be fairly similar to our comparison of the two-group designs in Chapter 10. Practical considerations become somewhat more important, however, in the multiple-group designs, so our conclusions will be somewhat different.

Choosing a Multiple-Group Design Again, your first consideration in choosing an experimental design should always be your experimental question. After you have decided on an experiment with one IV and three or more groups, you must determine whether you should use independent or correlated groups. If only one of those choices is viable, you have no further considerations to make. If, however, you could use either independent or correlated groups, you must make that decision before proceeding.

Control Issues As with the two-group designs discussed in Chapter 10, your decision to use the multiple-independent-groups design versus the multiple-correlated-groups design revolves around control issues. The multiple-independent-groups design uses the control technique of randomly assigning participants to groups. If you have a substantial number of research participants (at least 10 per group), you can be fairly confident that random assignment will create equal groups.

Multiple-correlated-groups designs use the control techniques of matching, repeated measures, or natural pairs to assure equality of groups and to reduce error variability. Recall the equation that represents the general formula for a statistical test:

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

Reducing the error variability in the denominator of the equation will result in a larger computed statistical value, thereby making it easier to reject the null hypothesis. We hope you remember from Chapter 10 that using a correlated-groups design reduces your degrees of freedom, which makes it somewhat more difficult to achieve statistical significance and reject the null hypothesis. The reduced error variability, however, typically more than offsets the loss of degrees of freedom. Correlated designs therefore often produce stronger tests for finding statistical significance.

Practical Considerations Matters of practicality become quite important when we contemplate using a multiple-correlated-groups design. Let's think about each type of correlated design in turn. If we intend to use *matched sets*, we must consider the potential difficulty of finding three (or more) participants to match on the extraneous variable we choose. Suppose

we conduct a learning experiment and thus wish to match our participants on IQ. How difficult will it be to find three, four, five, or more participants (depending on the number of levels we use) with the same IQ? If we cannot find enough to make a complete set of matches, then we cannot use those participants in our experiment. We may, therefore, lose potential research participants through the requirement of large matched sets. We may be limited in our use of *natural sets* by set size also. How much chance would you have of running an experiment on triplets, quadruplets, or quintuplets? For this reason, using animal littermates is probably the most common use of natural sets in multiple-group designs. When we use *repeated measures* in a multiple-group design, we are requiring each participant to be measured at least three times. This requirement necessitates more time for each participant or multiple trips to the laboratory, conditions the participants may not be willing to meet. We hope this message is clear: If you intend to use a multiple-correlated-groups design, plan it very carefully so that these basic practical considerations do not sabotage your experiment.

What about practical considerations in multiple-independent-groups designs? The multiple-independent-groups design is simpler than the correlated version. The practical factor you must take into account is the large number of research participants you will have to make random assignment feasible *and* to fill the multiple groups. If participants are not available in large numbers, you should consider using a correlated design.

Drawing a definite conclusion about running independent- versus correlated-multiple-group designs is not simple. The correlated designs have some statistical advantages, but they also require you to take into account several practical matters that may make using such a design difficult. Independent designs are simple to implement, but they force you to recruit or obtain many research participants to assure equality of your groups. The best advice we can provide is to remind you that each experiment presents you with unique problems, opportunities, and questions. You should be aware of the factors we have presented and to weigh them carefully in conjunction with your experimental question when you choose a specific research design for your experiment.

Variations on the Multiple-Group Design

In Chapter 10 we discussed two variations on the two-group design. Those same two variations are also possible with the multiple-group design.

Comparing Different Amounts of an IV This "variation" on the multiple-group design is not actually a variation at all; it is part of the basic design. Because the smallest possible multiple-group design would consist of three treatment groups, every multiple-group design must compare different amounts (or types) of an IV. Even if a multiple-group design has a control group, there are at least two different treatment groups in addition.

If we already know that a particular IV has an effect, then we can use a multiple-group design to help us define the limits of that effect. In this type of experiment we often add an important control in order to account for a possible **placebo effect**. For example, is it possible that some of the effects of coffee on our alertness are due to what we *expect* the coffee to do? If so, a proper control group would consist of people who drink decaffeinated coffee. These participants would be blind to the fact that their coffee does not contain caffeine. This group, without any caffeine, would show us whether coffee has any placebo effects.

Placebo effect An experimental effect caused by expectation or suggestion rather than the IV.

Dealing With Measured IVs All the research examples we have cited in this chapter deal with manipulated IVs. It is also possible to use measured IVs in a multiple-group design. In Chapter 4 you learned that the research we conduct with a measured rather than a manipulated IV is termed **ex post facto research**. Remember that we cannot randomly assign participants to groups in such research because they already belong to specific groups. Thus, the groups may be different at the beginning of the research. We cannot draw cause-and-effect relations from such an experiment because we do not directly control and manipulate the IV ourselves. Still, an ex post facto design can yield interesting information and, because the design does use some controls, we may be able to rule out some alternative explanations. Let's look at a student example of an ex post facto design with a measured IV.

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

Radha Dunham and her advisor Lonnie Yandell (2005), of Belmont University in Nashville, Tennessee, used an ex post facto approach in their study of students' feeling of self-efficacy about their drawing ability. They chose their participants from an advanced art class, a lower level art class, and a general psychology class to represent advanced, beginning, and non-art groups, respectively. Why does this experimental design fit the multiple-group format? Does it have one IV? Yes: level of art skill. Does it have three or more levels of that one IV? Yes: the advanced, beginning, and non-art groups. These levels were their measured IV—the researchers could not assign students to one of the art skill groups; they could only "measure" which class students were taking as their approximation of level of art skill.

Dunham and Yandell (2005) found that the non-art group showed the lowest level of art self-efficacy compared to both the advanced and beginning art groups. The beginning and advanced art groups did not differ in their art self-efficacy. Notice that the multiple-group design allowed Dunham and Yandell to detect a difference between one group versus each of the two other groups. This type of findings shows the advantage of the multiple-group design over the two-group design; two two-group experiments would have been necessary to obtain the results from this one multiple-group experiment. Remember these various types of differences because we will return to them in the "Statistical Analysis" section of the chapter, next.

REVIEW SUMMARY

1. Psychologists plan their experiments beforehand using an **experimental design**, which serves as a blueprint for the experiment.
2. You use the **multiple-group design** for experimental situations in which you have one **independent variable** that has three or more levels or conditions.
3. A multiple-group design may or may not use a **control group**. If there is a control group, there are at least two experimental groups in addition.
4. You form **independent groups** of research participants by **randomly assigning** them to treatment groups.
5. You form **correlated groups** of research participants by creating matched sets, using natural sets, or measuring each participant more than once (repeated measures).
6. **Multiple-correlated-groups designs** provide extra advantages for experimental control relative to **multiple-independent-groups designs**.

7. Practical considerations in dealing with research participants make the multiple-correlated-groups designs considerably more complicated than multiple-independent-groups designs.
8. Multiple-group designs exist primarily to allow comparisons of different amounts (or types) of IVs.
9. Measured IVs can be used in multiple-group designs, resulting in **ex post facto studies**.

Check Your Progress

1. Why is the two-group design the building block for the multiple-group design?
2. The simplest possible multiple-group design would have _____ IV(s) and _____ treatment group(s).
3. What advantage(s) can you see in using a multiple-group design rather than a two-group design?
4. Devise an experimental question that could be answered with a multiple-group design that you could not answer with a two-group design.
5. Why are matched sets, repeated measures, and natural sets all considered *correlated-groups* designs?
6. What is the real limit on the number of groups that can be included in a multiple-group design? What is the practical limit?
7. Make a list of the factors you would consider in choosing between a multiple-group design and a two-group design.
8. Correlated-groups designs are often advantageous to use because they _____.
9. Why are practical considerations of using a multiple-correlated-groups design more demanding than those when using a two-correlated-groups design or a multiple-independent-groups design?
10. If we wished to compare personality traits of firstborn, lastborn, and only children, what type of design would we use? Would this represent a true experiment or an ex post facto study? Why?

Statistical Analysis: What Do Your Data Show?

We will remind you from the previous chapter that experimental design and statistical analysis are intimately linked. You *must* go through the decision-making process we have outlined before you begin your experiment in order to avoid the possibility that you will run your experiment and collect your data only to find out that there is no statistical test that you can use to analyze your data.

Analyzing Multiple-Group Designs

In this chapter we have looked at designs that have one IV with three (or more) groups. In your introductory statistics course you probably learned that researchers analyze these

One-way ANOVA A statistical test used to analyze data from an experimental design with one independent variable that has three or more groups (levels).

Completely randomized ANOVA This one-way ANOVA uses independent groups of participants.

Repeated-measures ANOVA This one-way ANOVA uses correlated groups of participants.

multiple-group designs with the *analysis of variance* (ANOVA) statistical procedure. As you will see, we will also use ANOVA to analyze designs that include more than one IV (see Chapter 12); hence, there are different types of ANOVAs, and we need some way to distinguish among them. In this chapter we are looking at an ANOVA for one IV; researchers typically refer to this procedure as a **one-way ANOVA**.

You remember that we have considered both multiple-independent-groups and multiple-correlated-groups designs in this chapter. We need two different types of one-way ANOVA to analyze these two types of designs, just as we needed different *t* tests in Chapter 10. As you can see from Figure 11-1, when we assign our participants to multiple groups randomly, we will analyze our data with a one-way ANOVA for independent groups (also known as a **completely randomized ANOVA**). On the other hand, if we use matched sets, natural sets, or repeated measures, we will use a one-way ANOVA for correlated groups (also known as a **repeated-measures ANOVA**) to evaluate our data.

Planning Your Experiment

In Chapter 10 we featured the statistical analysis of data from an experiment designed to compare the response time of saleclerks as a function of their customers' clothing (also see Chapter 9). That example, of course, cannot serve as a data analysis example for this chapter because it represents a two-group design.

Suppose we have already conducted the sample experiment covered in Chapters 9 and 10. How could we conduct a similar experiment using a multiple-group design?

The *most* similar experiment would be one in which students in the introductory class dressed in three different types of clothing rather than just two. Suppose that we decide to investigate further because we found (in Chapter 10) that salesclerks responded more quickly to customers in dressy clothes than to those dressed in sloppy clothes. We decide to add an intermediate clothing group—we choose to add casual clothing as our third group. Again, we must consider the **operational definition** of our new IV group. We define casual clothing as slacks and shirts (e.g., khakis and polo shirts) for both male and female customers. We have 24 students in the class, so we have 8 students as "stimuli" in each of the three groups (one group for each type of clothing). We have the students go to the same store on the same day and randomly choose a department in which to browse. The store is large and employs many clerks, so there is no problem finding a different clerk for each student. This random choice will allow the salesclerks to be randomly assigned to the three groups (a requirement to create independent groups). An observer goes with the students to time unobtrusively the salesclerks' response time to each student, which is the dependent variable (DV). You can see the clerks' response times in Table 11-1. Let's discuss the basis behind the ANOVA procedure before we look at our statistical analyses.

Operational definition Defining the independent, dependent, and extraneous variables in terms of the operations needed to produce them.

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

TABLE 11-1 Salesclerks' Response Times (in Seconds) for Hypothetical Clothing Style Experiment

Clothing Styles		
Dressy	**Sloppy**	**Casual**
37	50	39
38	46	38
44	62	47
47	52	44
49	74	50
49	69	48
54	77	70
69	76	55
Mean = 48.38	Mean = 63.25	Mean = 48.88

Rationale of ANOVA

We expect that you learned something about ANOVA in your statistics course. We introduced a closely related concept in the Control Issues section in Chapter 10; you may wish to refer back to that section. You will remember that variability in your data can be divided into two sources: **between-groups variability** and **error variability** (also known as **within-groups variability**). The between-groups variability represents the variation in the DV that is due to the IV; the error variability is due to such factors as individual differences, errors in measurement, and extraneous variation. In other words, error variability refers to *any* variability in the data that is not a product of the IV. Look at Table 11-2, which is a slightly altered version of Table 11-1.

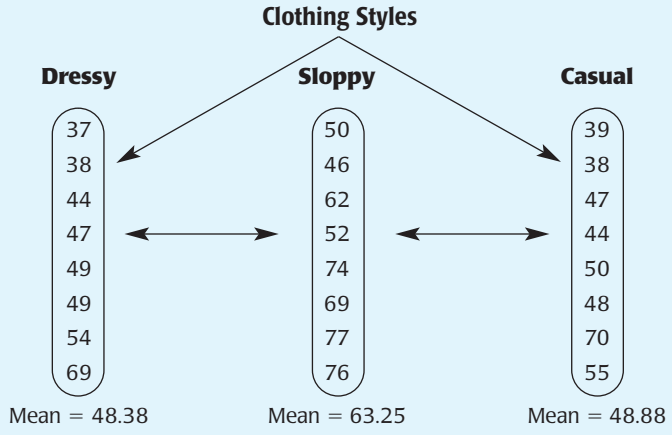
Error variability
Variability in DV scores that is due to factors other than the IV, such as individual differences, measurement error, and extraneous variation (also known as *within-groups variability*).

Within-groups variability
Another term for error variability.

What type of variability do you think is represented by the arrows in Table 11-2 (see next page)? What type of variability do you think is shown in the circled columns? If you have trouble with these questions, reread the previous paragraph.

The variability *between* the groups' response times represents the variability caused by the IV (the different types of clothing); therefore, the arrows represent the between-groups variability. If the times differ among the three groups of clerks, the clothes should be responsible for the difference (assuming we have controlled extraneous variables). On the other hand, error variability should occur among the participants *within* each particular group (thus its name, *within-groups variability*); this is the variability represented by the circled columns. One major source of within-groups variability is the differences within all the participants in a group—what we have labeled individual differences. Different people (or nonhuman animals) will score differently on the DV simply because they are different organisms.

TABLE 11-2 Salesclerks' Response Times (in Seconds) for Hypothetical Clothing Style Experiment



"Wait a minute," you may say. "What we have just described as within-groups variability—individual differences, measurement errors, extraneous variation—can occur between the groups just as easily as within the groups." This thought represents very good thinking on your part. Your point is correct and is well taken. Thus, we must change the formula that we reviewed just a few pages ago:

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

The fact that we can find error between our groups as well as within our groups forces us to alter this formula to the general formula shown below for ANOVA. The *F* symbol is used for ANOVA in honor of Sir Ronald A. Fisher (1890–1962), who developed the ANOVA (Spatz, 2001).

$$F = \frac{\text{variability due to IV} + \text{error variability}}{\text{error variability}}$$

If our IV has a strong treatment effect and creates much more variability than all the error variability, we should find that the numerator of this equation is considerably larger than the denominator (see Figure 11-4A). The result, then, would be a large *F* ratio. If, on the other hand, the IV has absolutely no effect, there would be no variability due to the IV, meaning we would add 0 for that factor in the equation. In such a case, our *F* ratio should be close to 1 because the error variability between groups should approximately equal the error variability within groups. This situation is depicted in Figure 11-4B.

The notion that has evolved for the ANOVA is that we are comparing the ratio of between-groups variability (variability caused by the IV) to within-groups variability. Thus, the *F* ratio is conceptualized (and computed) with the following formula:

$$F = \frac{\text{between-groups variability}}{\text{within-groups variability}}$$

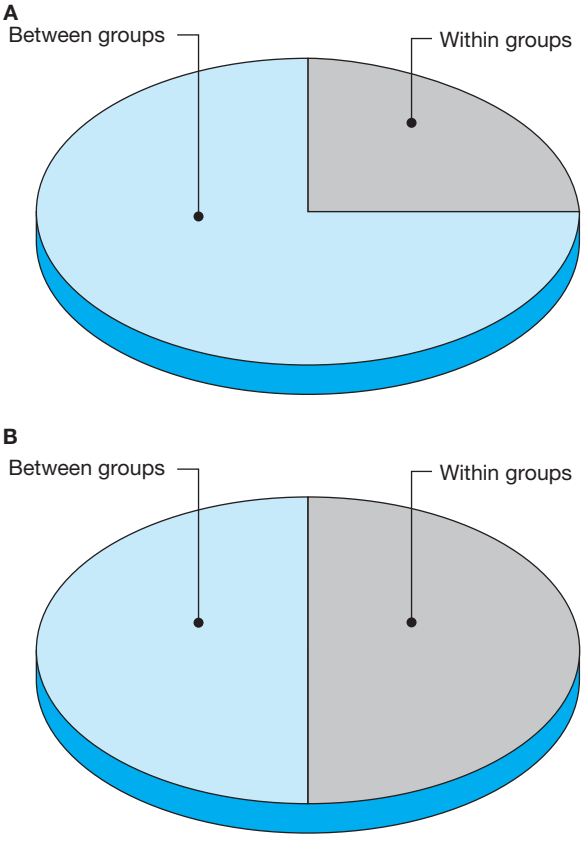


FIGURE 11-4 Possible Distributions of Variability in an Experiment. **A** depicts a large *F* ratio; **B** depicts an *F* ratio of 1.

A simple way to think of ANOVA is to realize that you are dividing the treatment effect by the error. When the IV has a significant effect on the DV, the *F* ratio will be large; when the IV has no effect or only a small effect, the *F* ratio will be small (near 1). You may wish to place a bookmark at this page—we will refer back to it shortly.

Interpretation: Making Sense of Your Statistics

With the addition of a third group, our experimental design has become slightly more complicated than the two-group design discussed in Chapter 10. As you will see, adding a third group (or more) creates an interesting statistical problem for us; we may have to compute an extra statistical test to explore significant findings. (Just in case you ever see such an analysis, one-way ANOVA can also be used for experiments with only two groups. We presented only the *t* test for such designs in the previous chapter to minimize overlap and possible confusion.)

Interpreting Computer Statistical Output

Once again we will look at generic computer output to give you experience with typical output so that you can better generalize your knowledge to the particular statistical package that is available to you. The results appear in Table 11-3.

One-Way ANOVA for Independent Samples We are examining results from a one-way ANOVA because we have one IV with three groups. We used the ANOVA for independent samples because we randomly assigned salesclerks to the three different clothing conditions. The DV scores represent the clerks' response times to differently attired customers.

As usual, we first look for information about descriptive statistics. You will find the descriptive statistics in the top portion of Table 11-3. Before going on, remember that we recommended that you make sure you have entered your data correctly in the computer by checking the means using a calculator. It will take only a few minutes but will spare you from using an incorrect set of results if you somehow goofed when you put the numbers in the computer. We can see that Group 1 (clerks responding to customers in dressy clothes) had a mean response time of 48.38 seconds, Group 2 (sloppy clothing) had a mean of 63.25 seconds, and Group 3 (casual clothes) responded in 48.88 seconds on the average. So, we do see numerical differences among these means, but we do not know whether the differences

TABLE 11-3 Computer Output for One-Way ANOVA for Independent Samples

GROUP	N	MEAN	STD DEV	STD ERR	95% CONF INT
1 (dressy)	8	48.38	10.11	3.57	39.92-56.83
2 (sloppy)	8	63.25	12.54	4.44	52.76-73.74
3 (casual)	8	48.88	10.20	3.61	40.34-57.41

ONEWAY ANOVA: RESPTIME by CLOTHING

SOURCE	SUM OF SQUARES	DF	MEAN SQUARES	F RATIO	PROB.
BETWEEN GROUPS	1141.75	2	570.88	4.71	.02
WITHIN GROUPS	2546.25	21	121.25		
TOTAL	3688.00	23			

POST HOC TEST: Tukey-HSD with significance level .05

* Indicates significant differences shown for pairings

Mean	CLOTHING	Grp 1	Grp 2	Grp 3
48.38	Grp 1			
63.25	Grp 2	*		*
48.88	Grp 3			

are large enough to be significant until we examine the inferential statistics. We also see the standard deviation and standard error (standard deviation divided by $\sqrt{n}$) for each group (the times for group 2 are more variable than those of the other two groups), as well as 95% confidence intervals. You may remember that confidence intervals provide a range of scores between which μ (the true population mean) should fall. Thus, we are 95% confident that the interval of 40.34 to 57.41 seconds contains the population mean for all clerks responding to customers in casual clothing.

The inferential statistics appear immediately below the descriptive statistics. We see the heading "ONEWAY ANOVA," which lets us know that we have actually computed a one-way ANOVA. The subheading shows us that we have analyzed the variable "RESPTIME" in relation to the "CLOTHING" variable. This label simply means that we have analyzed our DV (RESPTIME, the clerks' response times) by our IV (CLOTHING, the three styles of dress).

The output from ANOVA is typically referred to as a **source table**. In looking at the table, you will see "SOURCE" printed on the left side of the page. Source tables get their name because they isolate and highlight the different *sources* of variation in the data. In the one-way ANOVA table, you see two sources of variation: between groups and within groups.

Source table A table that contains the results of ANOVA. *Source* refers to the source of the different types of variation.

The two terms "between groups" and "within groups" refer to what?

Between groups is synonymous with our treatment (IV) effect, and *within groups* is our error variance. The **sum of squares**, the sum of the squared deviations around the mean, is used to represent the variability of the DV in the experiment (Kirk, 1968). We use ANOVA to divide (partition) the variability into its respective components, in this case between-groups and within-groups variability. In Table 11-3 you see that the total sum of squares (variability in the entire experiment) is 3688, which we partitioned into between-groups sum of squares (1141.75) and within-groups sum of squares (2546.25). The between-groups sum of squares added to the within-group sum of squares should always be equal to the total sum of squares (1141.75 + 2546.25 = 3688).

Sum of squares The amount of variability in the DV attributable to each source.

If we formed a ratio of the between-groups variability and the within-groups variability based on the sums of squares, we would obtain a ratio of less than 1. We cannot use the sums of squares for this ratio, however, because each sum of squares is based on a different number of deviations from the mean (Keppel, Saufley, & Tokunaga, 1992). Think about this idea for a moment: Only three groups can contribute to the between-groups variability, but many different participants can contribute to the within-groups variability. Thus, to put them on an equal footing, we have to transform our sums of squares to **mean squares**. We make this transformation by dividing each

Mean square The "averaged" variability for each source; computed by dividing each source's sum of squares by its degrees of freedom.

sum of squares by its respective degrees of freedom. Because we have three groups, our between-groups degrees of freedom are 2 (number of groups minus 1). Because we have 24 participants, our within-groups degrees of freedom are 21 (number of participants minus the number of groups). Our total degrees of freedom are equal to the total number of participants minus 1, or 23 in this case. As with the sums of squares, the between-groups degrees of freedom added to the within-groups degrees of freedom must equal the total degrees of freedom (2 + 21 = 23). Again, our mean squares are equal to each sum of squares divided by its degrees of freedom. Thus, our between-groups mean square is 570.88 (1141.75/2), and our within-groups mean square is 121.25 (2546.25/21).

Variance A single number that represents the total amount of variation in a distribution; also the square of the standard deviation, σ^2.

We should note at this point that a mean square is analogous to an estimate of the **variance**, which you may remember from statistics as the square of the standard deviation (σ^2). As soon as we have the mean squares, we can create our distribution of variation. Rather than drawing pie charts, like those shown in Figure 11-4, we compute an *F* ratio to compare the two sources of variation. Referring to the bookmark we advised you to use a few pages back, we find that the *F* ratio is equal to the between-groups variability divided by the within-groups variability. Because we are using mean squares as our estimates of variability, the equation for our *F* ratio becomes

$$F = \frac{\text{mean square between groups}}{\text{mean square within groups}}$$

Thus, our *F* ratio of 4.71, as shown in Table 11-3, was derived by dividing 570.9 by 121.3. This result means that the variability between our groups is almost five times larger than the variability within the groups. Or, perhaps more clearly, the variability caused by the IV is almost five times larger than the variability resulting from error. If we drew a pie chart for these results, it would look like Figure 11-5.

Finally, we come to the conclusion (or so we think!). Did the different clothing styles have a significant effect? Next to "F RATIO" in Table 11-3 you see the "PROB" entry: .02. This probability of chance of these data (if the null hypothesis is true) is certainly lower than .05, so we did find a significant difference. The difference in the response times among the three groups of salesclerks probably did not occur by chance. Although the computer printed the probability of

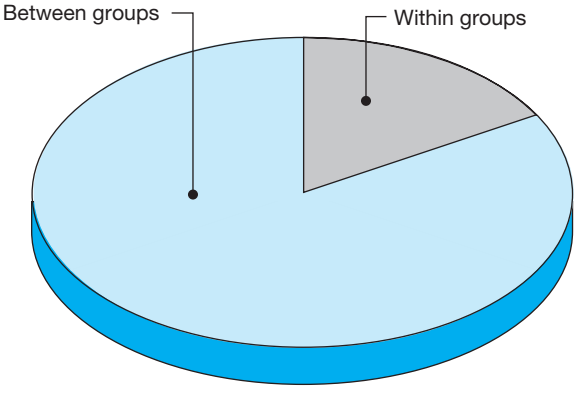


FIGURE 11-5 Distribution of Variability for Different Clothing Experiment.

chance for you, you should know how to use a printed *F* table just in case your computer program does not calculate probabilities. This activity is somewhat different from using a printed *t* table because in ANOVA we have two values for degrees of freedom. In this case our degrees of freedom are 2 (between groups, in the numerator) and 21 (within groups, in the denominator). When we look in the *F* table (see Table A-2), the column shows the numerator *df* and the row shows the denominator *df*. In this case you must find 2 in the numerator column and 21 on the denominator line and locate the intersection of those two numbers. At that point you will see 3.47 as the .05 cutoff and 5.78 as the .01 cutoff. Because our *F* value of 4.71 is between these two numbers, the probability that it could have occurred by chance is less than .05 but greater than .01. Thus, if you were using a table rather than the computer output, you would have written $p < .05$ *or* $.01 < p < .05$ (to be as specific as possible). Sherlock Holmes stated, "I could only say what was the balance of probability" (Doyle, 1927, p. 93).

With the two-group design, we would be finished with our computer output at this point and could go on to interpreting our statistics in words. With significant findings in a two-group design, we merely note the higher mean and conclude that it is significantly higher than the other mean. This decision procedure is not correct, however, with the multiple-group design when we find a significant difference because we have more than two means. We know there is significance among our means because of our significant *F* ratio, but which one(s) is (are) different from which one(s)? From a significant *F* ratio, we cannot tell.

Post hoc comparisons
Statistical comparisons made between group means after finding a significant *F* ratio.

To discern where the significance lies in a multiple-group experiment, we must conduct additional statistical tests known as **post hoc comparisons** (also known as *follow-up tests*). These tests allow us to determine which groups differ significantly from each other after we have determined that there is overall significance (by finding a significant *F* ratio). Many different post hoc tests exist, and there is much debate over these tests that is beyond the scope of this text. Simply remember that you will have to conduct post hoc tests if you find overall significance in a one-way ANOVA.

At the bottom of Table 11-3, you see the results of a post hoc test known as the Tukey HSD (an abbreviation for *honestly significant difference*). The Tukey test allows you to test all pair-wise comparisons, meaning that you can test the difference between all sets of two means (Keppel et al., 1992). In looking at Table 11-3, we see that Group 2 is significantly different at the .05 level from both Group 1 and Group 3 according to the Tukey test. This result means that the clerks took significantly longer to wait on sloppily dressed students (63.25 seconds) than either those in casual (48.88 seconds) or those in dressy clothes (48.38 seconds). No other groups differed significantly from each other, meaning that there was no statistical difference in the time it took clerks to wait on students dressed well or casually.

As in the previous chapter, we hope that you are learning general principles about computer printouts rather than specific words or terms for which you will blindly search. If you understand the general principles, interchanging *between groups* with *clothing* (or the name of some other IV) should not be problematical for you; different statistical programs may simply use different ways of getting at the same thing (much like having slightly different names for the same test). For example, don't be surprised to see the label *error* rather than *within groups*—both terms mean the same thing. The important conclusion is that given the same data, any two programs should find the same results.

Translating Statistics Into Words Let us remind you, as we did in Chapter 10, that the results of any statistical test are only as good as your experimental procedures. In other words, if you have conducted a sloppy experiment, your statistical results will be meaningless. When we draw the conclusion that our IV has caused a difference in the DV scores, we are assuming that we conducted a well-controlled experiment and removed extraneous variables from the scene. If you find that extraneous variables have confounded your experiment, you should not interpret your statistics because they are now meaningless. For the same reason, detectives must learn specific ways to collect evidence in the field. If they collect contaminated evidence, all the lab tests in the world cannot yield a definitive conclusion.

Based on our inferential statistics, we can conclude that the clothing customers wear is important because clerks took differing amounts of time to wait on customers depending on how they were dressed.

Although this conclusion is technically correct, it is a poor conclusion. Why? How would you revise this conclusion to make it better?

This conclusion is poor because it is incomplete. Reread the sentence and decide what you can learn from it. All you know is that students who wore some type of clothing were waited on more quickly than other students who wore some other clothing. Thus, you know that clothes can make a difference, but you don't know which type of clothing led to faster responses. To write a good conclusion, we must go back to our inferential statistics, particularly the post hoc tests. In Table 11-3 we find that the students wearing dressy clothes were waited on in 48.38 seconds, students wearing sloppy clothes received attention in 63.25 seconds, and students wearing casual clothes got help in 48.88 seconds. The significant *F* ratio lets us know that there is significance *somewhere* among those means. The Tukey post hoc comparison tests informed us that the differences between Group 2 and both Groups 1 and 3 were significant. To interpret this difference, we must examine our descriptive statistics. When we examine the means, we are able to conclude that students in shabby clothes got help significantly more slowly than the students in dressy or casual clothes. *No other mean differences were significant.*

We must determine how to communicate our statistical findings in APA format. We will use a combination of words and numbers. There are many different ways to write this set of findings in an experimental report. Here is one example:

> The effect of different clothing on salesclerks' response time was significant, $F(2, 21) = 4.71$, $p = .02$. The proportion of variance accounted for by the clothing effect (η^2) was .31. Tukey tests indicated ($p < .05$) that clerks waiting on customers dressed in sloppy clothes ($M = 63.25$, $SD = 11.73$) responded more slowly than clerks waiting on customers in dressy ($M = 48.38$, $SD = 9.46$) or casual clothes ($M = 48.88$, $SD = 9.55$). The response times of clerks waiting on customers in dressy and casual clothes did not differ from each other.

The words alone should convey the meaning of our results. Could someone with no statistical background read and understand these sentences if we removed the numbers? We think so. The inferential test results explain our findings to readers with a statistical background. The descriptive statistics allow the reader to observe exactly how the groups performed and how variable that performance was. The effect size information reported here, η^2 (eta squared), is similar to r^2 because it tells you the proportion of variance in the DV (response times) accounted for by

the IV (clothing). (An easy way to calculate η^2 is to divide the between-groups sum of squares by the total sum of squares.) The reader has an expectation about what information will be given because we write our results in this standard APA format. You will find this type of communication in results sections in experimental articles. As you read more results sections, this type of communication will become familiar to you.

One-Way ANOVA for Correlated Samples Now we will look at the one-way ANOVA for correlated samples. The sample experiment about clothing and clerks' response times we have used so far in the chapter fits the multiple-group design for independent samples and thus is not appropriate to analyze with the one-way ANOVA for correlated samples.

How could you modify the experiment concerning salesclerks' reactions to customers' style of dress so that it used correlated groups rather than independent groups?

To be correct, you should have proposed the use of matched sets, natural sets, or repeated measures in your modified experiment. The best choices would involve matched sets or repeated measures; we don't think that natural sets is a feasible choice in this situation—you're not using littermates, and finding sets of triplets who are all salesclerks is most unlikely! If you choose matched sets, you must decide on a matching variable. It is difficult to know what variable on which you should match salesclerks that would be related to their reactions to different clothing styles. Matched sets would not be a good choice for forming correlated groups.

Imagine that you conduct your experiment at a small store that employs only a few salesclerks. In order to have enough data points, you decide to have each salesclerk respond to a customer in each of the three clothing groups. Because you would measure each clerk's response times to all three styles of dress, you would control possible individual differences between the clerks. (Another scenario that might lead you to use repeated measures would occur if you decided that it is likely that some variable in salesclerks, such as attitude, might affect their response times to customers in different types of clothing. Using repeated measures would allow you essentially to cancel out differences between different clerks, because each clerk would wait on a customer in each clothing group.)

You are now ready to begin the experiment. The students each dress in one of the three styles and enter the store. Because we are using repeated measures, we know for certain that the clerks in the three groups are equated (because they are the same clerks in each group). Given this hypothetical example, the scores in Table 11-1 would now represent sets of response times from eight salesclerks. (Remember that in the real world, it is not legitimate to analyze the same data with two different statistical tests. This is a textbook, certainly not the real world, and we are doing this as an example only.)

You can see the results for the one-way ANOVA for correlated samples in Table 11-4. As usual, we are first interested in examining descriptive statistics. The descriptive output is shown at the top of Table 11-4. As you can see, we obtain the mean, standard deviation, sample size, and 95% confidence interval for each group. The descriptive statistics for the three groups match what we have previously seen in Table 11-3, which is certainly logical. Although we are now using a correlated-samples analysis, nothing has changed about the samples themselves. So, we see the same means, standard deviations, and confidence intervals that we have seen before.

TABLE 11-4 Computer Output for One-Way ANOVA for Correlated Samples

```
  GROUP          N      MEAN      STD DEV      STD ERR      95% CONF INT

1 (dressy)       8     48.38       10.11         3.57        39.92-56.83
2 (sloppy)       8     63.25       12.54         4.44        52.76-73.74
3 (casual)       8     48.88       10.20         3.61        40.34-57.41

ONEWAY ANOVA: RESPTIME by CLOTHING (CORR SAMP)

SOURCE              SUM OF      DF       MEAN       F RATIO      PROB.
                    SQUARES              SQUARES

CLOTHING           1141.75       2      570.88       19.71       .000
SUBJECTS           2140.65       7      305.81       10.56       .000
WITHIN CELLS        405.59      14       28.97
TOTAL              3688.00      23

POST HOC TEST: Tukey-HSD with significance level .01

* Indicates significant differences shown for pairings

                                G      G      G
                                r      r      r
                                p      p      p
                                1      2      3

Mean       CLOTHING

48.38        Grp 1
63.25        Grp 2              *             *
48.88        Grp 3
```

The other information that we see in Table 11-4 is our ANOVA source table. Once again, the entries in this particular source table vary slightly from the tables we've looked at earlier. Although you may begin to believe that this is some sinister plot hatched just to confuse you, you have to focus on the basic information, remembering that terms are used slightly differently in different situations. Here, our three sources of variance are labeled "within cells," "subjects," and "clothing." Because you know that we are comparing different types of clothing as our IV, it should be clear that "clothing" represents the effect of our IV and "within cells" represents our source of error variation (refer back to Table 11-2 to see our within-cell variation pictorially represented by the circles). "Subjects," of course, represents the variability between different salesclerks. When we examine the source table, we find that the *F* ratio for the comparison of the clerks' response times to different clothing styles is 19.71, with 2 (numerator) and 14 (denominator) degrees of freedom, which results in a probability of chance of .000 according to the computer.

This situation illustrates one of our pet peeves with computerized statistical programs. When you studied distributions in your statistics class, what did you learn about the tails of those distributions? We hope you learned that the tails of distributions are **asymptotic**; that is, the tails extend into infinity and never touch the baseline. This fact means that the probability of a statistic is *never* .000. No matter how large the statistic gets, there is always some small amount of probability under the tails of the distribution. Unfortunately, people who design statistics software either

Asymptotic Refers to tails of distributions that approach the baseline but never touch the baseline.

have a limited number of columns to work with *or* they don't think about this issue, so they have the computer print a probability of .000, implying that there is no uncertainty. In light of this problem, we advise you to list $p < .001$ if you ever find such a result on your computer printout.

Pardon the digression, but you know how pet peeves are! Back to the statistics. The overall effect of the clothing is significant, which leads us to wonder which clothing styles differ from each other. This source table looks different from the one in Table 11-3 because it shows the effects of two IVs: CLOTHING and SUBJECTS. Although the SUBJECTS effect is significant, it does not tell us anything very important: We simply learn that there were significant differences between the eight salesclerks' response times. In other words, we found individual differences between the salesclerks. This effect is an expected one and is not profound. Typically, you would ignore this effect. However, the SUBJECTS effect *is* important statistically. If you compare Tables 11-3 and 11-4, you will see that the correlated samples ANOVA has taken the SUBJECTS variability (mean square) out of the WITHIN CELLS (or error) variability compared to the WITHIN GROUPS (or error) term in the independent-samples ANOVA. This difference demonstrates the power of the correlated-samples analysis to reduce variability in the error term and to create a larger *F* ratio.

As with the multiple-group design for independent samples, we used a Tukey test for post hoc comparisons. Again, we found that Group 2 (sloppy clothing) was significantly different ($p < .01$) from both Group 1 (dress clothes) and Group 3 (casual clothes); however, Groups 1 and 3 did not perform significantly differently from each other. Notice that our significant differences are at the .01 level rather than .05 as with the independent-samples case. This change is another indication of the increased power of the correlated-samples analysis.

Translating Statistics Into Words Our experimental logic is no different from that for the independent-samples ANOVA. The only difference is that we used a somewhat more stringent control procedure in this design—we used repeated measures with our participants rather than assigning them to groups randomly.

Our conclusions should combine our numbers with words to give the reader a clear indication of what we found. Remember to include information both about any difference that was found and the directionality of the difference.

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

Although the conclusion for the correlated-samples test is similar to that for the independent-groups test, it is different in some important ways. We hope you figured out those important differences. Here's a sample conclusion:

> The effect of three different clothing styles on clerks' response times was significant, $F(2, 14) = 19.71$, $p < .001$. The proportion of variance accounted for by the clothing effect (η^2) was .74. Tukey tests showed ($p < .01$) that clerks took longer to respond to customers dressed in sloppy clothes ($M = 63.25$, $SD = 11.73$) than to either customers in dressy clothes ($M = 48.38$, $SD = 9.46$) or customers in casual clothes ($M = 48.88$, $SD = 9.55$). Response times did not differ between the clerks waiting on customers in dressy or casual clothing.

Did your conclusion look something like this? Remember, the exact wording may not necessarily match—the important thing is that you cover all the critical details.

There are five important differences between this conclusion and the conclusion drawn for the independent-groups ANOVA. Can you find them?

The *first* difference comes in the degrees of freedom. There are fewer degrees of freedom for the error term in the correlated-groups ANOVA (WITHIN CELLS) than for the independent-samples case (WITHIN GROUPS). *Second*, the *F* value for the correlated-groups test is larger than for the independent-groups test. The larger *F* value is a result of reducing the variability in the denominator of the *F* equation. This difference in *F* values leads to the *third* difference, which is the probability of chance. Despite the fact that there are fewer degrees of freedom for the correlated-samples case, its probability of chance is lower (smaller). *Fourth*, the proportion of variance accounted for by the clothing effect (η^2) was considerably larger. *Fifth*, the post hoc tests show a lower (smaller) probability of chance in the correlated-groups situation.

These last three differences most clearly show the advantage of a correlated-groups design. Because using repeated measures reduced some of the error variability, the probability of the difference coming about by chance is smaller than it was in the independent-samples case. Thus, the conclusion from the correlated-groups design yields the clearer finding (reducing the chance of a Type I error). We cannot promise that correlated-groups designs will always allow you to find a clearer difference than independent-groups designs; however, we can tell you that correlated-groups designs do increase your odds of detecting smaller significant differences because such designs reduce error variance.

The Continuing Research Problem

In Chapters 9 and 10 we began our continuing research project by looking at clerks' response times as a function of how customers were dressed. Clerks' times were significantly higher when they waited on customers in sloppy clothing than when they waited on well-dressed customers. Because of this result, we decided to pursue this line of research further and, in this chapter, compared the effects of three different styles of clothing to each other. On the basis of our results, we can state that salespeople wait on customers in dressy or casual clothing more quickly than they wait on customers in sloppy clothing.

Is our research project complete at this point? As you might have realized, we could compare an endless number of styles of dress. This research problem could go on forever. In all seriousness, you might have wondered about the effects of other possible IVs on salesclerks' response times. As we begin to ask more complicated questions, we must move on to more complex designs to handle those questions. In Chapter 12 we will be able to continue our research problem with an experiment using more than one IV at a time.

Let's review the logical steps we took in conducting this experiment. Refer back to Figure 11-1 to take note of our experimental design questions.

1. After conducting a preliminary experiment (Chapter 10) and determining that salesclerks waited on well-dressed customers more quickly, we decided to test further the effects of different clothing (IV) on clerks' response times (DV).

2. We chose to test only one IV (clothing) because our research is still preliminary.
3. We tested three different styles of dress because they seemed to be valid ways for customers to dress.

4a. With access to many salesclerks, we used random assignment to the three groups and, thus, a multiple-independent-groups design. We used a one-way ANOVA for independent groups and found that clerks responded more quickly to customers in dressy or casual clothes than to customers in sloppy clothes.

4b. With smaller numbers of clerks, we chose to use repeated measures. Thus, we used a multiple-within-group design and a one-way ANOVA for correlated groups. Clerks responded to sloppily dressed customers more slowly than to well-dressed or casually dressed customers.

5. We concluded (hypothetically) that customers should not dress in sloppy clothes if they desire to get helped quickly in a store.

REVIEW SUMMARY

1. When your experimental design consists of one IV with three or more groups and you have randomly assigned participants to groups, the proper statistical analysis is a one-way ANOVA for independent groups (**completely randomized ANOVA**).
2. When your experimental design has one IV with more than two groups and you have used matched sets, natural sets, or repeated measures, you should analyze your data with a one-way ANOVA for correlated groups (**repeated-measures ANOVA**).
3. ANOVA partitions the variability in your DV into **between-groups variability** (caused by the IV) and **within-groups variability** (resulting from sources of error). We then compute a ratio between these two sources of variation known as the *F* ratio.
4. ANOVA results are typically shown in a **source table**, which lists each source of variance and displays the *F* ratio for the effect of the IV.
5. A significant *F* ratio merely indicates that there is a significant difference somewhere among your various groups. **Post hoc comparisons** are necessary to determine which groups differ from each other.
6. Using APA format for our statistical results allows us to convey our findings in both words and numbers in a clear and concise manner.
7. Previous experiments often lead to further questions and new experiments. The multiple-group design is an ideal design to follow up on the results from a two-group experiment.

Check Your Progress

1. Suppose you wish to compare the ACT or SAT scores of the freshman, sophomore, junior, and senior classes at your school to determine whether differences exist among those students. Draw a block diagram of this design. What design and statistical test would you use to conduct this research?

2. You wonder whether students who take the ACT or SAT three times are able to improve their scores significantly. You select a sample of such students and obtain their three scores. What type of experimental design does this question represent? Draw a block diagram of it. What statistical test would you use to analyze the data?
3. When we look at our *F* ratio and its probability in a multiple-group design, why can't we examine the descriptive statistics directly to reach a conclusion about our experiment?
4. The variability that is due to our IV is termed the _____ variance, whereas the variability caused by individual differences and error is the _____ variance.
5. Suppose you conducted the experiment summarized in Question 2 and found the following statistics: $F(2, 24) = 4.07$, $p < .05$. On the basis of this information, what could you conclude?
6. What additional information do you need in Question 5 to draw a full and complete conclusion?
7. In the continuing research problem from this chapter, why was it important to have the (hypothetical) knowledge from the similar study in Chapter 10?
8. You decide to test how people's moods vary by the four seasons. What type of experimental design would you use for this research project? Why?
9. You choose to test people's preferences for fast-food hamburgers, and you have McDonald's, Burger King, Wendy's, and White Castle franchises in your town. What type of experimental design would you use for this research project? Why?

Key Terms

Experimental design, 231
Independent variable, 232
Principle of parsimony, 233
Levels, 233
Treatment groups, 233
Independent groups, 235
Correlated groups, 235
Random assignment, 235
Control procedure, 235
Confounded experiment, 235
Matching variable, 236
Placebo effect, 241
Ex post facto research, 242
One-way ANOVA, 244
Completely randomized ANOVA, 244
Repeated-measures ANOVA, 244
Operational definition, 244
Between-groups variability, 245
Error variability, 245
Within-groups variability, 245
Source table, 249
Sum of squares, 249
Mean square, 249
Variance, 250
Post hoc comparisons, 251
Asymptotic, 254

Looking Ahead

In this chapter we furthered our knowledge about research design and how it fits with particular experimental questions. Specifically, we looked at an extension of the basic building-block design by using one IV and three or more groups. In the next chapter we will make a significant alteration in our basic design by adding a second IV. This expanded design will give us the ability to ask much more sophisticated questions about behavior because most behaviors are affected by more than one variable at a time.