Experimental Design II: Factorial Designs

8

PREVIEW & CHAPTER OBJECTIVES

Chapter 7 introduced you to some basic experimental designs—those involving a single independent variable, with two or more levels of that variable being compared. The beauty of these designs is their simplicity, but human behavior is immensely complex. Thus, researchers often prefer to examine more than a single factor in their studies. To do this methodologically, the next logical step is to increase the number of independent variables being examined. When a study includes more than a single independent variable, the result is called a *factorial design*, the focus of this chapter. When you complete this chapter, you should be able to:

- Describe factorial designs using a standardized notation system (2 × 2, 3 × 5, etc.).
- Place data accurately into a factorial matrix, and calculate row and column means.
- Understand what is meant by a main effect, and know how to determine if one exists.
- Understand what is meant by an interaction effect, and know how to determine if one exists.
- Know how to interpret interactions and know the presence of an interaction sometimes lessens or eliminates the relevance of a main effect.
- Describe the research design of Jenkins and Dallenbach's (1924) famous study on sleep and memory and explain why their results could be considered an interaction.
- Identify the varieties of factorials corresponding to the single-factor designs of Chapter 7 (independent groups, matched groups, ex post facto, repeated measures).
- Identify a mixed factorial design, and understand why counterbalancing is not always used in such a design.
- Identify a P × E factorial design and understand what is meant when such a design produces main effects and interactions.
- Distinguish mixed P × E factorial from simple P × E factorial designs.
- Calculate the number of subjects needed to complete each type of factorial design.
- Know how to be an ethically responsible experimenter.

As you have worked your way through this research methods course, you have probably noticed that experimental psychologists seem to have a language of their own. They talk about operationalizing constructs, rejecting null hypotheses, and

eliminating confounds, and when they talk about regression, they are not discussing Freud. You haven't seen anything yet. After mastering this chapter, you will be able to say things like this: "It was a two by three mixed factorial that produced one main effect for the repeated measures variable plus an interaction." Let's start with the basics.

Essentials of Factorial Designs

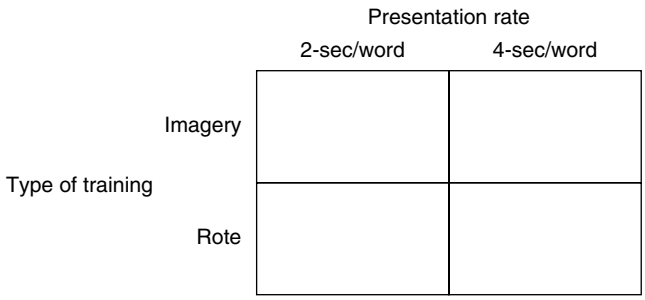
Suppose you are interested in memory and wish to find out if recall can be improved by training people to use visual imagery while memorizing a list of words. You could create a simple two-group experiment in which some people are trained to use visual imagery techniques while memorizing the words ("create a mental image of each word") and others are told to use rote repetition ("just repeat the words over and over to yourself"). Suppose you also wonder about how memory is affected by how quickly words are presented, that is, a word list's presentation rate. Again, you could do a simple two-group study in which some participants see the lists at the rate of 2 seconds per word, others at 4 seconds per word. With a factorial design, *both* of these studies can be done as part of the same experiment, allowing the researcher to examine the effects of both repetition and presentation rate, and to determine if these factors combine to affect memory.

By definition, a **factorial design** involves any study with more than one independent variable (recall from Chapter 7 that the terms *independent variable* and *factor* mean the same thing). In principle, factorial designs could involve dozens of independent variables, but in practice these designs usually involve two or three factors, sometimes four.

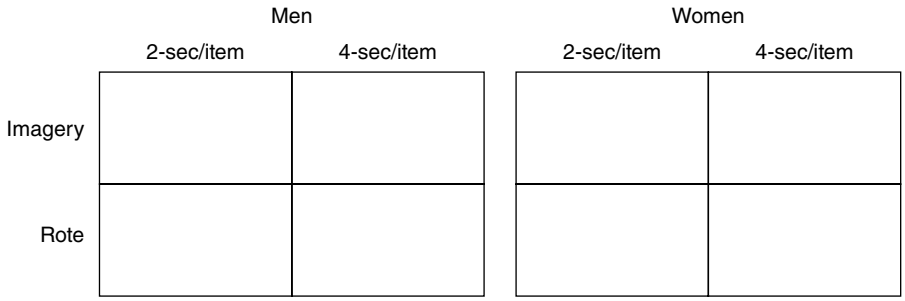
Identifying Factorial Designs

A factorial design is described with a numbering system that simultaneously identifies the number of independent variables and the number of levels of each variable. Each digit in the system represents an independent variable, and the numerical value of each digit indicates the number of levels of each independent variable. Thus, a 2 × 3 (read this as "two by three") factorial design has two independent variables; the first has two levels and the second has three. A more complex design, a 3 × 4 × 5 factorial, has three independent variables with three, four, and five levels, respectively. The hypothetical memory study we just described would be a 2 × 2 design, with two levels of the "type of training" independent variable (imagery and rote repetition) and two levels of the "presentation rate" independent variable (2 and 4 seconds per item).

The total number of conditions to be tested in a factorial study can be identified by looking at all possible combinations of the levels of each independent variable. In our hypothetical memory study, this produces a display called a **factorial matrix**, which looks like this:



Before going on, note this carefully: Up to this point in the book, we have been using the concepts "conditions of the experiment" and "levels of the independent variable" as if they meant the same thing. These concepts indeed are interchangeable in single-factor experiments (i.e., one independent variable). In factorial designs, however, this is no longer the case. In all experimental designs, the term *levels* refers to the number of levels of any one independent variable. In factorial designs, the term *conditions* equals the number of cells in a matrix like the one you just examined. Hence, the 2 × 2 memory study has *two* independent variables, each with *two* levels. It has *four* conditions, however, one for each of the four cells. The number of conditions in any factorial design can be determined by multiplying the numbers in the notation system. Thus, a 3 × 3 design has nine conditions; a 2 × 5 has ten conditions, and a 2 × 2 × 2 has eight conditions. Incidentally, although the use of a factorial matrix is a bit awkward when there are three independent variables, as in the 2 × 2 × 2 just mentioned, a matrix can be drawn. Suppose our memory study added gender as a third factor, in addition to presentation rate and type of training. The factorial matrix could look like this:



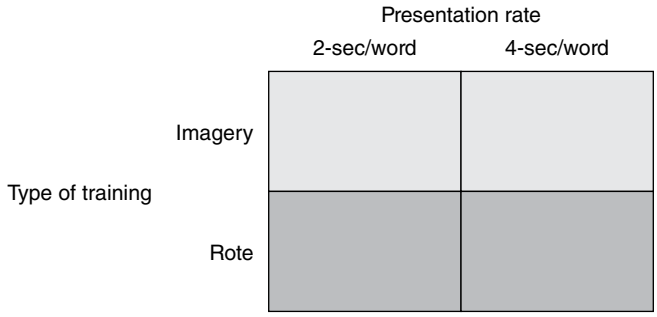
Outcomes—Main Effects and Interactions

In factorial studies, two kinds of results occur: main effects and interactions. *Main effects* refer to the overall influence of each of the independent variables, and *interactions* examine whether the variables combine to form a more complex result. Let's look at each in more detail.

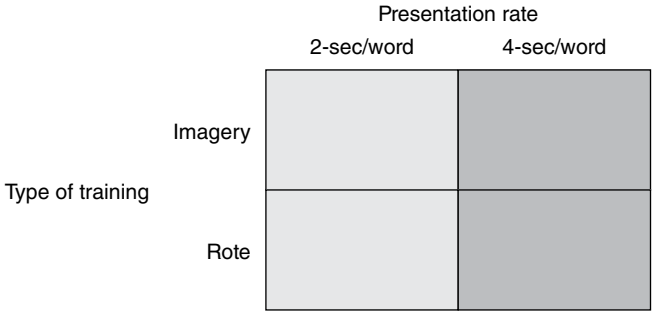
Main Effects

In the memory experiment we've been using as a model, the researcher is interested in the effects of two independent variables: type of training and presentation rate. In factorial designs, the term **main effect** is used to describe the overall effect of a single independent variable. Specifically, a main effect is the difference between the means of the levels of any one independent variable. So, in a study with two independent variables, such as a 2 × 2 factorial, there can be at most two significant main effects. Determining the main effect of one factor involves combining all of the data for each of the levels of that factor. In our hypothetical memory study, this can be illustrated as follows. The main effect of type of training is determined by combining the data for participants trained to use imagery (for both presentation rates combined) and comparing it to all of the data for participants using rote repetition. Hence, all of the information in the lightly shaded cells

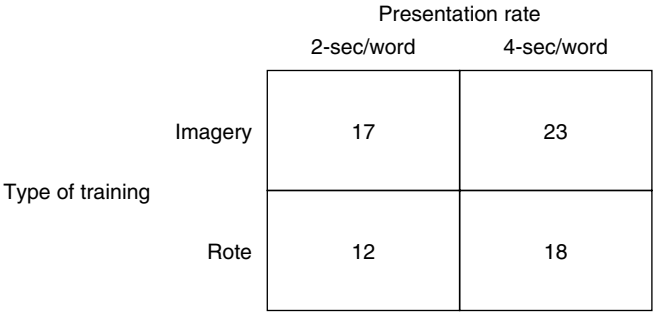
(imagery) of the following matrix would be combined and compared with the combined data in the more heavily shaded cells (rote):



Similarly, the main effect of presentation rate is determined by combining the data for everyone presented the words at a 2-second rate and comparing that with the data from those presented the words at a 4-second rate. In the following matrix, the effect of presentation rate would be evaluated by comparing all of the information in the lightly shaded cells (2-sec/item) with all of the data in the more heavily shaded cells (4-sec/item):



Let's consider hypothetical data for a memory experiment like the example we've been using. Assume 25 subjects in each condition (i.e., each cell of the matrix). Their task is to memorize a list of 30 words. The average number of words recalled for each of the four conditions might look like this:



Does imagery training produce better recall than rote repetition? That is, is there a main effect of type of training? The way to find out is to compare all of the "imagery" data with all of the

"rote" data. Specifically, this involves calculating *row means*. The "imagery" row mean is **20** words [(17 + 23)/2 = 40/2 = 20], and the "rote" row mean is **15** words [(12 + 18)/2 = 30/2 = 15]. When asking if training type is a main effect, the question is: "Is the difference between the row means of 20 and 15 statistically significant or due to chance?"

In the same fashion, calculating *column means* allows us to see if presentation rate is a main effect. For the 2 sec/item column, the mean is **14.5** words; it is **20.5** words for the 4 sec/item row (you should check this). Putting all of this together yields this outcome:

		Presentation rate		
		2-sec/word	4-sec/word	**Row means**
Type of training	Imagery	17	23	**20.0**
	Rote	12	18	**15.0**
Column means		**14.5**	**20.5**	

For these data, it appears that imagery improves memory (20 > 15) and that recall is higher if the words are presented at a slower rate (20.5 > 14.5). That is, there seem to be two main effects here (of course, it takes an analysis of variance (ANOVA) to make a judgment about whether the differences are significant statistically or due to chance—more on *factorial ANOVAs* later in this chapter). For a real example of a study that produced two main effects, consider this example of the so-called closing time effect.

Research Example 19—Main Effects

We don't know how often country music produces empirical questions that intrigue research psychologists, but one example is a song produced by Mickey Gilley in 1975, "The Girls All Get Prettier at Closing Time," which includes the lyrics: "Ain't it funny, ain't it strange, the way a man's opinion changes, when he starts to face that lonely night." The song suggests that, as the night wears on, men in bars, desperate to find a companion for the night, lower their "attractiveness" threshold—the same woman who only seemed moderately attractive at 10:00 p.m. becomes more eye-catching as closing time looms. Yes, pathetic. Nonetheless, several researchers have ventured courageously into bars and clubs on the edges of campuses to test this "closing time" concept, and to determine whether it applies to both men *and* women who are searching for. . .whatever. One interesting example is a study by Gladue and Delaney (1990). In addition to resulting in two main effects (gender and time), the study also illustrates several other methodological points.

Gladue and Delaney's (1990) study took place over a 3-month period at a large bar that included a dance floor and had the reputation as being a place where "one had a high probability of meeting someone for subsequent romantic. . .activities" (p. 380). The researchers recruited 58 male and 43 female patrons, who made attractiveness ratings of a set of photographs of men and women and also made global ratings of the overall attractiveness of the people who happened to be at the bar—"Overall, how attractive would you rate the men/women in the bar right now?" (p. 381). The ratings, made at 9:00 p.m., 10:30 p.m., and 12:00 midnight, were on a 10-point

scale, with 10 being the most attractive. For the global ratings, here is the factorial matrix for the results (means estimated from the graph in Figure 8.1):

	Time period			
	9:00	10:30	12:00	**Row means**
Men rating women	5.6	6.4	6.6	**6.2**
Women rating men	4.8	5.2	5.6	**5.2**
Column means	**5.2**	**5.8**	**6.1**	

Both main effects were significant (remember that main effects are determined by examining row and column means). In this case, the average ratings increased for both men *and* women as the night wore on (column means → 5.2 < 5.8 < 6.1). Also, women were generally more discerning—they rated men lower than men rated women during all three periods combined (row means → 5.2 < 6.2). Figure 8.1 is Gladue and Delaney's (1990) bar graph of the same data. Note the use of error bars.

You might be thinking that one potential confound in the study was alcohol use. As one drinks more during the course of an evening, others might come to be seen as more attractive. So alcohol consumption could be confounded with the time of the attractiveness rating. Gladue and Delaney (1990) dealt with this by measuring alcohol intake and eliminated the problem by finding no overall relationship between intake amount and the attractiveness ratings. Another problem was that the global ratings of attractiveness were relatively crude measures, open to a number of interpretations. For instance, the actual people in the bar at the three times were probably different (people come and go during the evening), so the ratings at the three times might reflect actual attractiveness differences in those at the bar. To account for this problem, Gladue and Delaney also asked their subjects to rate photos of college-age female and male students. The same photos were used for all three periods. These photos had been pretested for levels of attractiveness, and photos with

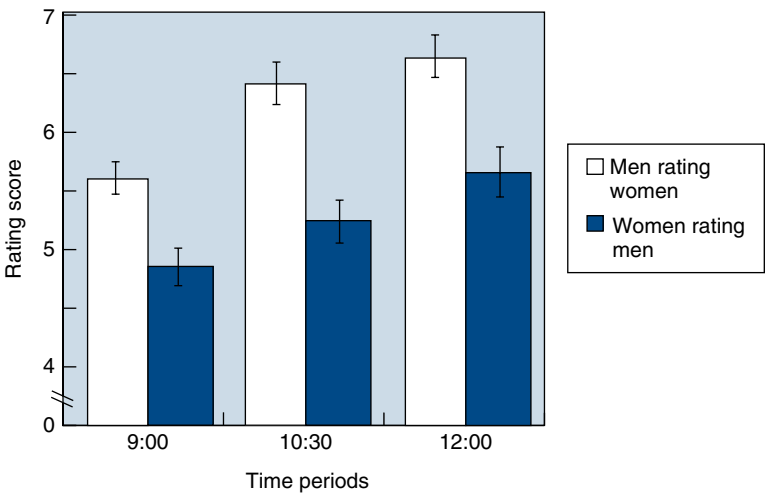


FIGURE 8.1
Main Effects—attractiveness ratings over the course of an evening for men and women rating each other, from the "closing time" study by Gladue and Delaney (1990).

moderate degrees of attractiveness had been chosen by the researchers. Why moderate? The researchers wished to avoid a problem first mentioned in Chapter 5—*ceiling effects* and *floor effects*. That is, they wanted to use photos for which changes in ratings, either up or down, would be likely to occur during the three periods. The ratings of the photos produced a more subtle effect than the global ratings. As the time passed, a closing time effect occurred for men (confirming Mickey Gilley, perhaps), but it did *not* occur for women; their ratings stayed about the same across all three time periods. This outcome is known as an *interaction*, our next topic.

SELF TEST

8.1

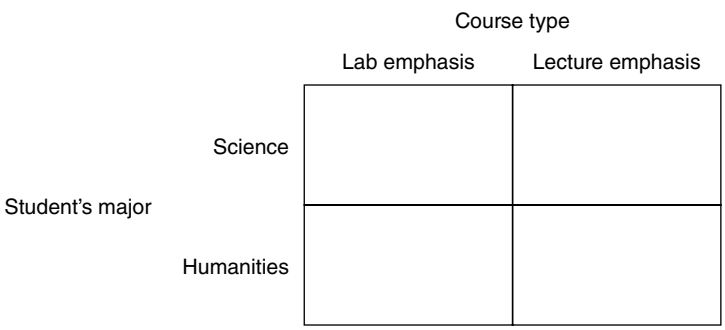
1. A 2 × 3 × 4 factorial design has (a) how many IVs, (b) how many levels of each IV, and (c) how many total conditions?
2. What is the basic definition of a main effect?
3. A memory study with a 2 (type of instruction) × 2 (presentation rate) factorial, like the example used at the start of the chapter, has these results (DV = words recalled):

   Imagery/2-sec rate = 20 words
   Imagery/4-sec rate = 20 words
   Rote/2-sec rate = 12 words
   Rote/4-sec rate = 12 words

   Are there any apparent main effects here? If so, for which factor? or both? Calculate the row and column means.

Interactions

Main effects are important outcomes in factorial designs, but the distinct advantage of factorials over single-factor designs lies in their potential to show interactive effects. In a factorial design, an **interaction** is said to occur when the effect of one independent variable depends on the level of another independent variable. This is a moderately difficult concept to grasp, but it is of immense importance because interactions often provide the most interesting results in a factorial study. In fact, interactions sometimes render main effects irrelevant. To start, consider a simple example. Suppose we hypothesize that an introductory psychology course is best taught as a laboratory self-discovery course rather than as a straight lecture course, but we also wonder if this is generally true or true only for certain kinds of students. Perhaps science majors would especially benefit from the laboratory approach. To test the idea, we need to compare a lab with a lecture version of introductory psychology, but we also need to compare types of students, perhaps science majors and humanities majors. This calls for a 2 × 2 design that looks like this:



In a study like this, the dependent variable would be some measure of learning; let's use a score from 1 to 100 on a standardized test of knowledge of general psychology, given during final exam week. Suppose these results occurred:

Student's major	Course type: Lab emphasis	Course type: Lecture emphasis
Science	80	70
Humanities	70	80

Are there any main effects here? No—all of the row and column means are the same: 75. So did anything at all happen in this study? Yes—something clearly happened. Specifically, the science students did better in the lab course than in the lecture course, but the humanities students did better in the lecture course than in the lab course. Or, to put it in terms of the definition of an interaction, the effect of one variable (course type) depended on the level of the other variable (major). Hence, even if no main effects occur, an interaction can occur and produce an interesting outcome.

This teaching example also highlights the distinct advantage of factorial designs over single-factor designs. Suppose you completed the study as a single-factor, two-level design, comparing lab with lecture versions of introductory psychology. You would probably use a matched group design, with student GPA and perhaps major as matching variables. In effect, you might end up with the same people who were in the factorial example. However, by running it as a single-factor design, your results would be:

Lab course : 75 Lecture course : 75

and you might conclude it doesn't matter whether introductory psychology includes a lab or not. With the factorial design, however, you know the lab indeed matters, *but only for certain types of students*. In short, factorial designs can be more informative than single-factor designs. To further illustrate the concept of interactions, in this case one that also failed to find main effects, consider the outcome of the following study.

Research Example 20—An Interaction with No Main Effects

Considerable research indicates that people remember information best if they are in the same general environment or context where they learned the information in the first place. A typical design is a 2 × 2 factorial, with the independent variables being the situation when the material is studied and the situation when the material is recalled. A nice example, and one that has clear relevance for students, is a study conducted by Veronica Dark and a group of her undergraduate students (Grant et al., 1998).

Grant et al.'s (1998) study originated from a concern that students often study under conditions quite different from the test-taking environment: They often study in a noisy environment but then take their tests in a quiet room. So, in the experiment, participants were asked to study a two-page article on psychoimmunology. Half of the participants studied the article while listening (over headphones) to background noise from a tape made during a busy lunchtime in a cafeteria (no distinct voices, but a "general conversational hum that was intermixed with the sounds produced by movement of chairs and dishes"—p. 619); the remaining participants studied the

same article without any background noise (but with the headphones on—can you see why they used headphones in this second group?). After a short break, all of the participants were tested on the material (with short answer and multiple-choice questions), also in either a noisy or quiet environment. This 2 × 2 independent groups design yielded the following four conditions:

1. silent study — silent recall
2. noisy study — noisy recall
3. silent study — noisy recall
4. noisy study — silent recall

The results for both the short answer and the multiple-choice tests showed the same general pattern; here are the mean scores for the multiple-choice results (max score = 16):

	Study condition: Silent	Study condition: Noisy	**Row means**
Silent during recall	14.3	12.7	**12.8**
Noisy during recall	12.7	14.3	**13.5**
Column means	**12.8**	**13.5**	

This outcome is similar to the pattern found in the hypothetical study about ways of teaching introductory psychology to science and humanities students. Row and column means were close (12.8 and 13.5), and they were not significantly different from each other. So there were no main effects. But examining the four individual cell means shows an interaction clearly occurred. When the students studied the essay in peace and quiet, they recalled well in the quiet context (14.3) but not so well in the noisy context (12.7); when they studied in noisy conditions, they recalled poorly in a quiet context (12.7) but did well when recalling when it was noisy (14.3). That is, learning was best when the study context matched the recall context. To put it in interaction language, the effect of one factor (where they recalled) depended on the level of the other factor (where they studied). Figure 8.2 presents the data in bar graph form.

This study had important limitations. First, there were few subjects (total of 39). Second, there were several experimenters, all undergraduate students of uncertain training and consistency as experimenters. Nonetheless, the predicted interaction occurred, one that is similar to the results of other studies with the same basic design (e.g., Godden & Baddeley, 1975), perhaps an indication of an effect strong enough to overcome methodological weaknesses.

Can you see the relevance of this for your life as a student? First, unlike much of the research on this context effect, which typically uses lists of words for study, Grant et al. (1998) used study material similar to the kind you would encounter as a student—text information to be comprehended. So the study has a certain amount of *ecological validity* (Chapter 5). Second, although you might conclude from these data that it doesn't matter whether you study in a quiet or noisy environment (just be sure to take the test in the same kind of environment), a fact of academic life is that tests are taken in quiet rooms. Unless you can convince your professors to let you take tests with your iPod going full blast (don't count on it), this study suggests it is clearly to your advantage to study for exams in a quiet place.

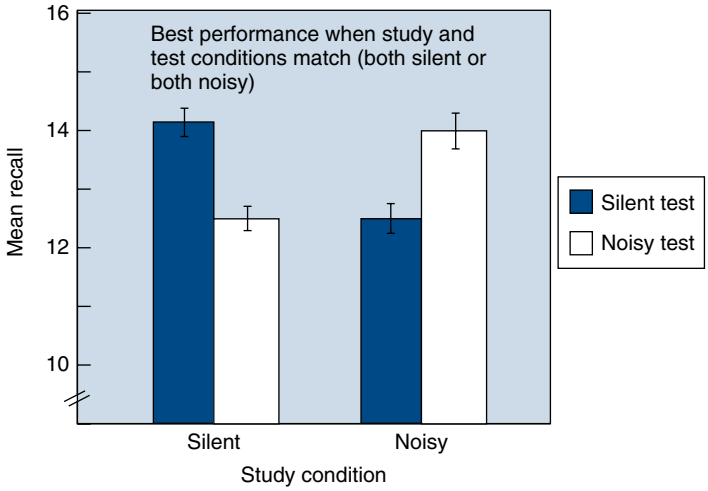


FIGURE 8.2
Bar graph showing an interaction between study and recall conditions (constructed from data in Grant et al., 1998).

Interactions Sometimes Trump Main Effects

In the opening paragraph describing interactions, you might have noticed a comment about interactions sometimes making main effects irrelevant. This frequently occurs in factorial designs for a specific type of interaction. A good example will make the point. Research Example 9 (Chapter 6) was designed to illustrate the use of a double blind procedure, but it also produced a significant interaction and two significant but irrelevant main effects. As you recall, the study examined the effect of caffeine on the memory of elderly subjects who were self-described "morning people." When tested in the morning, they did equally well whether taking caffeine in their coffee or having decaf. In the late afternoon, however, they did well with caffeine, but poorly with decaf. Here are the data in factorial matrix form:

	Time of day: Morning	Time of day: Afternoon	**Row means**
Caffeinated coffee	11.8	11.7	**11.8**
Decaffeinated coffee	11.0	8.9	**10.0**
Column means	**11.4**	**10.3**	

In this experiment, both main effects were statistically significant. Overall, recall was better for caffeine than for decaf (11.8 > 10.0) and recall was also better for morning sessions than afternoon sessions (11.4 > 10.3). If you carefully examine the four cell means, however, you can see performance was about the same for three of the cells, and declined only for the cell with the 8.9 in it—decaf in the afternoon. You can see this effect even more clearly in Figure 8.3: Three of the bars are essentially the same height, while the fourth is much lower.

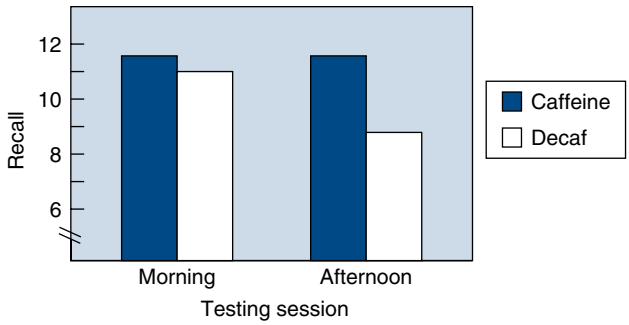


FIGURE 8.3
When an interaction renders main effects meaningless (from Ryan, Hatfield, & Hofstetter, 2002).

The only really important finding here is that recall declined in the afternoon for elderly adults drinking decaf—if they drank caffeinated coffee in the afternoon, they did as well as they had in the morning. Thus, you do not get a true picture of the key result if you report that, overall, caffeine produced better memory than decaf (11.8 > 10.0). In fact, caffeine's only advantage was in the afternoon; in the morning, whether subjects drank caffeine or decaf did not matter. Similarly, emphasizing the second main effect, that recall was generally better in the morning than in the afternoon (11.4 > 10.3), also gives a false impression of the key result. Recall was only better in the morning when decaf was consumed; when caffeine was used, morning or afternoon didn't matter. In short, for this kind of outcome, the interaction is the only important result. The tip-off that you are dealing with the kind of interaction where main effects do not matter is a graph like Figure 8.3, where three of the bars (or points on a line graph) are essentially the same, and a fourth bar (or point) is very different in height.

Combinations of Main Effects and Interactions

The experiment on studying and recalling with or without background noise (Research Example 20) illustrates one type of outcome in a factorial design (an interaction, but no main effects), but many patterns of results could occur. In a simple 2 × 2 design, for instance, there are eight possibilities:

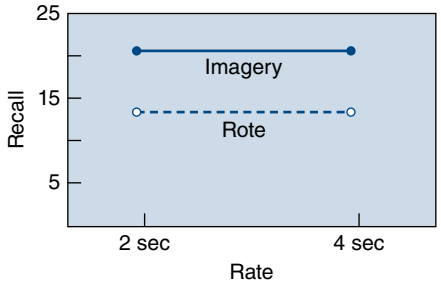
1. a main effect for the first factor only
2. a main effect for the second factor only
3. main effects for both factors; no interaction
4. a main effect for the first factor plus an interaction
5. a main effect for the second factor plus an interaction
6. main effects for both factors plus an interaction
7. an interaction only, no main effects
8. no main effects, no interaction

Let's briefly consider several of these outcomes in the context of the earlier hypothetical experiment on imagery training and presentation rate. For each of the following examples, we have created data that might result from the study on the effects of imagery instructions and presentation rate on memory for a 30-word list, translated the data into a line graph, and verbally

described the results. We haven't tried to create all of the eight possibilities listed above; rather, the following examples illustrate outcomes likely to occur in this type of study.

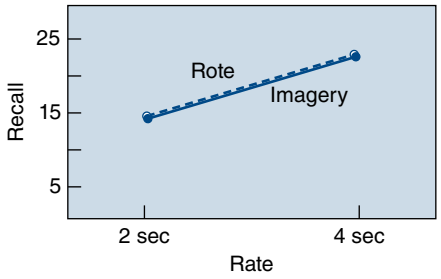
1. Imagery training improves recall, regardless of presentation rate; presentation rate doesn't affect recall. That is, there is a main effect for type of training factor—imagery (22) is better than rote (14). There is no main effect for presentation rate, however—the 2-sec rate (18) equals the 4-sec rate (18).

	2 sec	4 sec	Overall
Imagery	22	22	22
Rote	14	14	14
Overall	18	18	



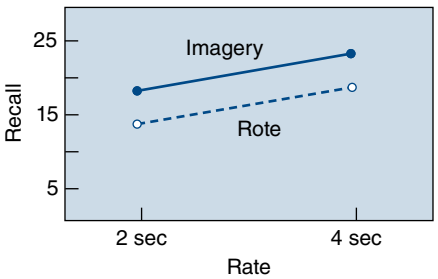
2. Recall is better with slower rates of presentation, but the imagery training was not effective in improving recall. That is, there is a main effect for the presentation rate factor—recall was better at 4-sec/item (22) than at 2-sec/item (14). But there was no main effect for type of training—imagery (18) was the same as rote (18).

	2 sec	4 sec	Overall
Imagery	14	22	18
Rote	14	22	18
Overall	14	22	



3. Recall is better with slower rates of presentation (20 > 16); in addition, the imagery training was effective in improving recall (20 >16). In this case, main effects for both factors occur. This is the outcome most likely to occur if you actually completed this study.

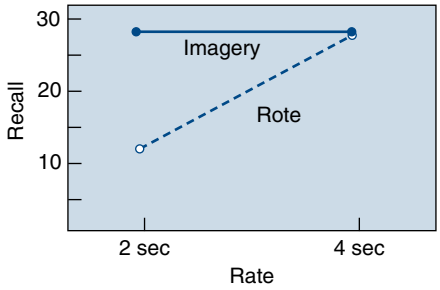
	2 sec	4 sec	Overall
Imagery	18	22	20
Rote	14	18	16
Overall	16	20	



4. At the 2-sec presentation rate, imagery training clearly improves recall (i.e., from 12 to 28); however, at the 4-sec rate, recall is almost perfect (28 = 28), regardless of how subjects are trained. Another way to say this is that when using imagery, presentation rate doesn't matter, but it does matter when using rote repetition. In short, there is an interaction between type of training and presentation rate. In this case, the interaction may have been influenced

by a *ceiling effect*, a result in which the scores for different conditions are all so close to the maximum (30 words in this example) that no difference could occur. Here, the imagery group recalls nearly all the words, regardless of presentation rate. To test for the presence of a ceiling effect, you could replicate the study with 50-item word lists and see if performance improves for the imagery/4-sec group.

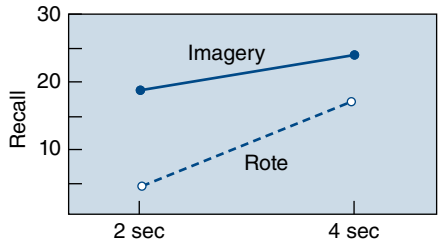
	2 sec	4 sec	Overall
Imagery	28	28	28
Rote	12	28	20
Overall	20	28	



You may be wondering about the obvious main effects that occur in this example. Surely the row (20 and 28) and column (also 20 and 28) means indicate significant overall effects for both factors. Technically, yes, the analysis probably would yield statistically significant main effects in this example, but this just illustrates again that interactions can trump main effects when the results are interpreted—the same point just made about the decaf-in-the-afternoon study. For the hypothetical memory study, the main effects are not meaningful; the statement that imagery yields a general improvement in recall is not really accurate. Rather, it only seems to improve recall at the faster presentation rate. Likewise, concluding that 4 seconds per item produces better recall than 2 seconds per item is misleading—it is only true for the rote groups. Hence, the interaction is the key finding here.

5. This is not to say that main effects never matter when an interaction exists, however. Consider this last example:

	2 sec	4 sec	Overall
Imagery	19	23	21
Rote	5	15	10
Overall	12	19	



In this case, imagery training generally improves recall (i.e., there's a main effect for type of training: 21 > 10). Also, a slower presentation rate improves recall for both groups (i.e., a main effect for presentation rate also: 19 > 12). Both of these outcomes are worth reporting. Imagery works better than rote at *both* presentation rates (19 > 5 and 23 > 15), and the 4-sec rate works better than the 2-sec rate for *both* types of training (23 > 19 and 15 > 5). What the interaction shows is that slowing the presentation rate improves recall somewhat for the imagery group (23 is a bit better than 19), but slowing the rate improves recall considerably for the rote rehearsal group (15 is a lot better than 5). Another way of describing the interaction is to say that at the fast rate, the imagery training is especially effective (19 is a lot better than 5—a difference of 14 items on the memory test). At the slower rate, imagery training still yields better recall, but not by as much as at the fast rate (23 is somewhat better than 15—a difference of just 8 items on the memory test).

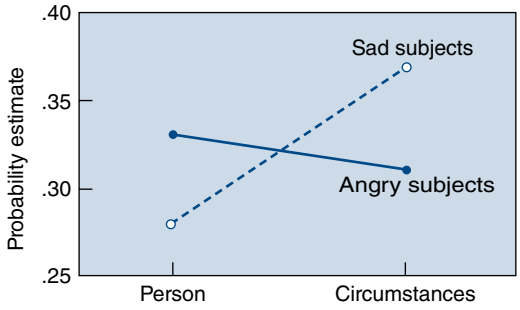


FIGURE 8.4
Using a line graph to highlight an interaction (from Keltner, Ellsworth, & Edwards, 1993).

From examining these graphs, you might have noticed a standard feature of interactions. In general, if the lines on the graph are parallel, then no interaction is present. If the lines are nonparallel, however, an interaction probably exists. Of course, this is only a general guideline. Whether an interaction exists (in essence, whether the lines are sufficiently nonparallel) is a statistical decision, to be determined by an ANOVA.

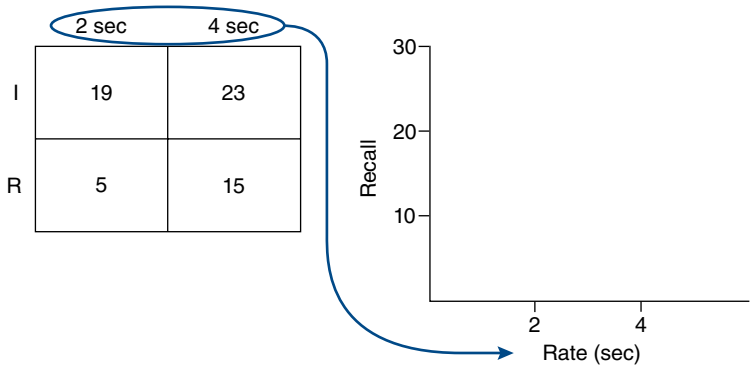
Identifying interactions by examining whether lines are parallel or not is easier with line graphs than with bar graphs. Hence, the guideline mentioned in Chapter 7 about line graphs being used only with within-subjects factors is sometimes ignored by researchers if the key finding is an interaction. For example, a study by Keltner, Ellsworth, and Edwards (1993) showed that when participants were asked to estimate the likelihood of a bad event (e.g., a car accident) occurring, there was an interaction between the emotion they experienced during the experiment and whether the hypothetical event was said to be caused by a person or by circumstances. When participants were feeling sad, they believed events produced by circumstances (e.g., wet roads) were more likely to occur than events produced by individual actions (e.g., poor driving). When participants were angry, however, the opposite happened—they believed events caused by individuals were more likely. As you can see from Figure 8.4, a line graph was used in the published study even though the *X*-axis uses a discrete variable. Keltner et al. (1993) probably wanted to show the interaction as clearly as possible, so they ignored the guideline about discrete variables. To repeat a point made earlier, when presenting any data, the overriding concern is to make one's hard-earned results as clear as possible to the reader.

Creating Graphs for the Results of Factorial Designs

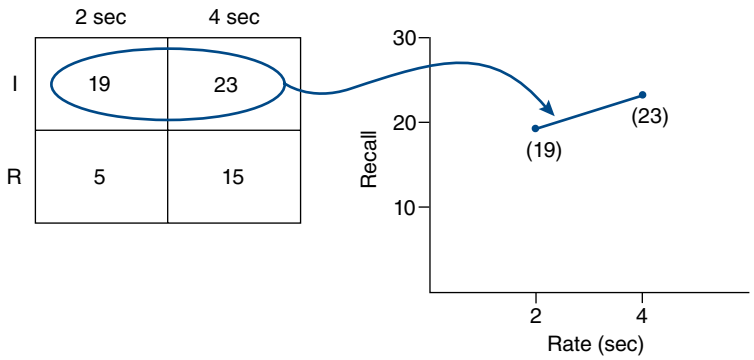
Whether it's the line or the bar version, students sometimes find it difficult to create graphs when studies use factorial designs. In single-factor designs, the process is easy; there is only a single independent variable, so there is no question about what will appear on the *X*-axis. With factorials, however, the situation is more complicated. A 2 × 2, for example, has two independent variables, but a graph has only one *X*-axis. How does one proceed?

This problem can be solved in a number of ways, but here is a simple and fool-proof system. Let's use Example 5 of the hypothetical imagery training and presentation rate study we have just worked through—the one with two main effects and an interaction. Creating the graph on the right from the matrix on the left can be accomplished as follows:

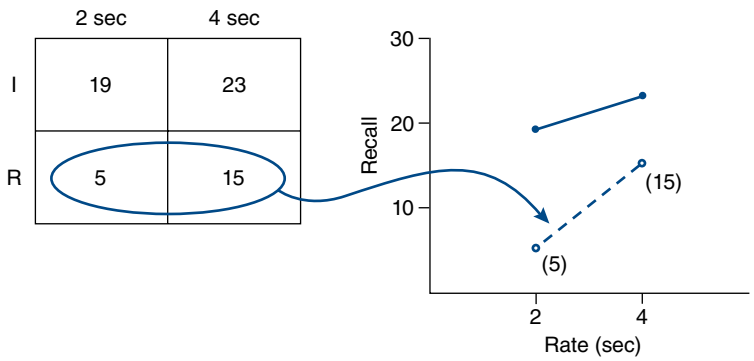
Step 1. Make the independent variable in the columns of the matrix the label for the *X*-axis. Think of this visually. The "2 sec" and "4 sec" are horizontal on the matrix—keep them horizontal on the graph and just slide them down to the *X*-axis. Label the *Y*-axis with the dependent variable.



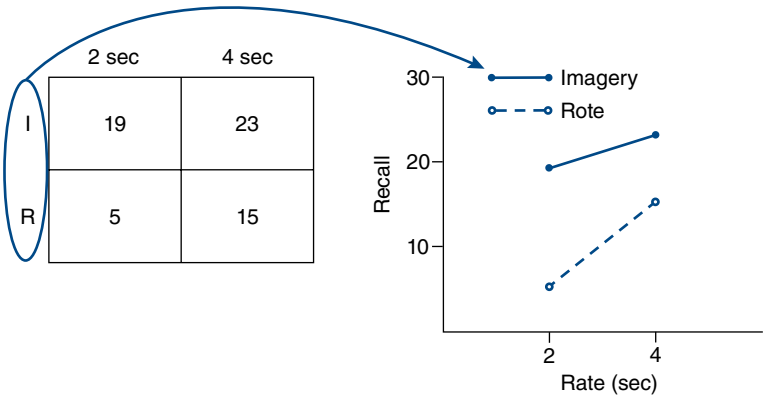
Step 2. Move the means from the top row of the matrix directly to the graph. Just as the "19" is on the left and the "23" is on the right side of the matrix, they wind up on the same left and right sides of the graph.



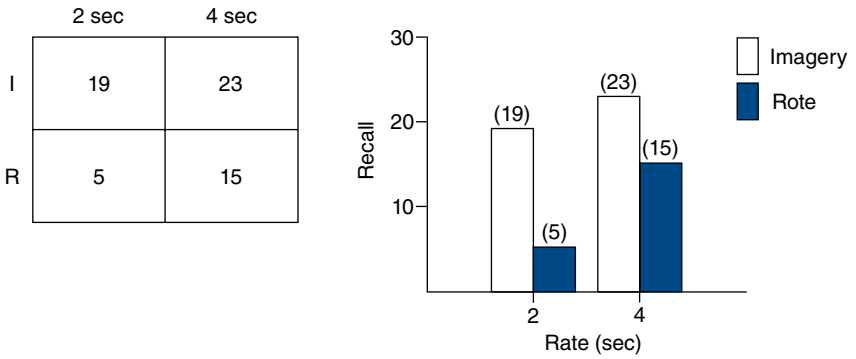
Step 3. Do the same with the means from the bottom row of the matrix ("5" and "15").



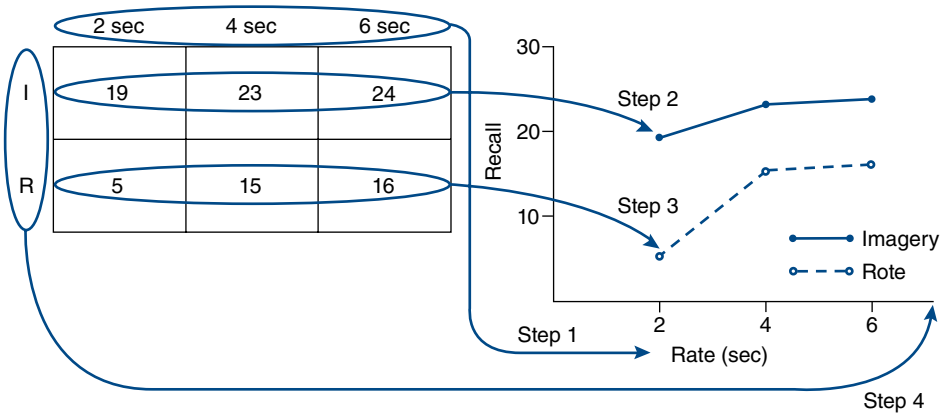
Step 4. Create a legend that identifies the second independent variable. Again think visually—"imagery" is above "rote" in the matrix and in the legend.



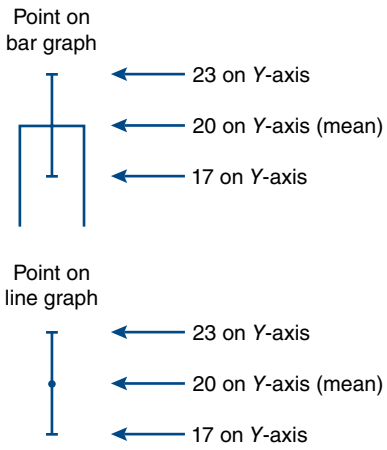
Note that if you wish to create a bar graph, the same basic process applies. Each of the points in the line graph turns into the top line of a bar.



Suppose the study added a third level to the presentation rate factor—6 sec per item, for instance. Here's how the graph building process would proceed:



You learned in Chapter 6 that in recent years it has become standard practice to include *error bars* on graphs. The points on a line graph and the tops of the bars on a bar graph are the mean scores; error bars tell you about the amount of variability in a set of scores. Error bars can be in the form of standard deviations, standard errors (an estimate of the population standard deviation, based on sample data), or confidence intervals (when SPSS does a graph, this is the default choice). Suppose you had a set of scores for which the mean was 20 and the standard deviation was 3. Here's how the error bars would look on both a bar graph and a line graph (for real examples in this chapter, look at Figures 8.8 and 8.10).



Before we turn to a system for categorizing types of factorial designs, you should read Box 8.1. It describes one of psychology's most famous experiments, a classic study supporting the idea that between the time you last study for an exam and the time you take the exam, you should be sleeping. It was completed in the early 1920s, when the term *factorial design* had not yet been invented and when analysis of variance, the statistical tool most frequently used to analyze factorials, was just starting to be conceptualized. Yet the study illustrates the kind of thinking that leads to factorial designs: the desire to examine more than one independent variable at the same time.

BOX 8.1 CLASSIC STUDIES—To Sleep, Perchance to Recall

Although the term *factorial design* and the statistical tools to analyze factorials were not used widely until after World War II, attempts to study more than one variable at a time occurred well before then. A classic example is a study by Jenkins and Dallenbach (1924) that still appears in many general psychology books as the standard example of retroactive interference (RI), or the tendency for memory to be hindered if other mental activities intervene between the time of study and the time of recall. In essence, the study was a 2 × 4 repeated-measures factorial design. The "2" was whether or not activities intervened between learning and a recall test, and the "4" referred to four retention intervals; recall was tested 1, 2, 4, or 8 hours after initial learning. What made the study interesting (and, eventually, famous) was the first factor. Participants spent the time between study and recall either awake and doing normal student behaviors, or asleep in Cornell's psychology lab. The prediction, that being asleep would produce less RI, and therefore better recall, was supported.

(continued)

BOX 8.1 (CONTINUED)

A close examination of the study illustrates some of the attributes of typical 1920s-era research and also shows that experimenters were just as careful about issues of methodological control then as they are now. As you will learn in Chapter 12, research in psychology's early years often featured very few participants. Compared with modern memory research, which uses many participants and summarizes data statistically, early studies were more likely to include just one, two, or three participants and report extensive data for each—additional participants served the purpose of replication. This happened in Jenkins and Dallenbach's (1924) study; there were just two subjects (referred to as *Observers* or *Os*, another typical convention of the time), both seniors at Cornell. When using small numbers of participants, researchers tried to get as much out of them as they could, and the result in this case was what we would call a repeated-measures study today. That is, both students contributed data to all eight cells of the 2 × 4 design, with each student learning and recalling lists eight times in each of the eight conditions—a total of 64 trials. If you are beginning to think the study was a major undertaking for the two Cornell seniors, you're right. During the study, the two students and Jenkins, who served as experimenter, "lived in the laboratory during the course of the experiments" (p. 606) in a simulated dorm room, and the study lasted from April 14, 1923, to June 7. Imagine giving up your last month and a half of college to science!

As good researchers, Jenkins and Dallenbach (1924) were concerned about control, and they used many of the procedures you've been learning about. For instance, they used 10-item lists of nonsense syllables, and the subjects read them aloud during the study trials until one perfect recitation occurred (i.e., their operational definition of learning). They took measures to ensure a consistent pronunciation of the syllables, and they used counterbalancing to avoid sequence effects in the presentation of the different retention intervals—"[t]he time-intervals between learning and reproduction were varied at haphazard" (p. 607). For the "awake" condition, the students learned their lists between 8 and 10 in the morning, then went about their normal business as students, and then returned to the lab for recall after 1, 2, 4, or 8 hours. For the "asleep" condition, lists were studied between 11:30 at night and 1:00 in the morning. Students then went to bed, and were awakened for recall by Jenkins 1, 2, 4, or 8 hours later. There was one potential confound in the study: On the awake trials, the students were told when to return to the lab for recall (i.e., they knew the retention interval), but during the asleep trials, students did not know when they would be awakened. Jenkins and Dallenbach were aware of the problem, considered alternatives, but decided their procedure was adequate.

The results? Figure 8.5 reproduces their original graph showing the data for each student. Each data point is an average of the eight trials for each condition of the study. Several things are clear. First, both students ("H" and "Mc") behaved similarly. Second, and this was the big finding, there was a big advantage for recall after sleeping, compared with recall after being awake. Third, there is the hint of an interaction. As Jenkins and Dallenbach (1924) described it: "The curves of

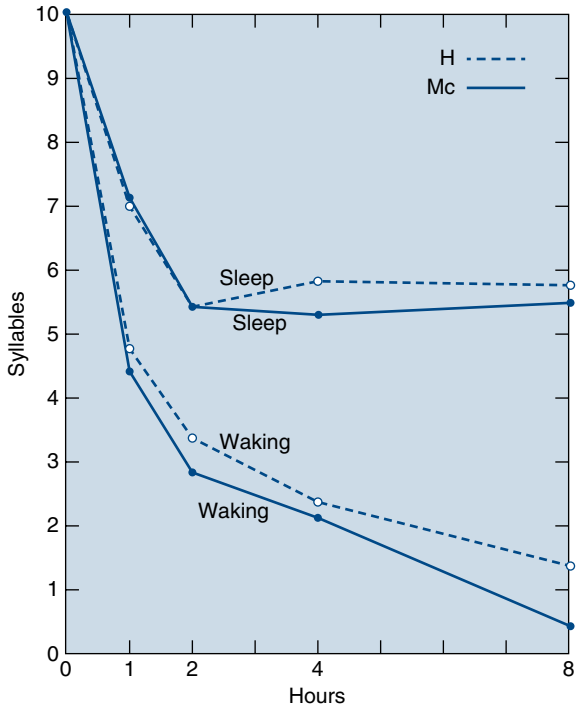


FIGURE 8.5
The Jenkins and Dallenbach study on retroactive interference, showing data for both of the Cornell students who participated, L. R. Hodell (H) and J. S. McGrew (Mc). Keep in mind that the study was completed long before an ethics code would have deleted the participants' names (from Jenkins & Dallenbach, 1924).

the waking experiments take the familiar form: a sharp decline which becomes progressively flatter. The form of the curves of the sleep experiments, however, is very different: after a small initial decline, the curves flatten and a high and constant level is thenceforth maintained" (p. 610).

One other intriguing outcome of the study is never reported in textbook accounts. As the experiment progressed, it became increasingly difficult for Jenkins to wake up the students. It was also hard for Jenkins to "know when they were awake. The Os would leave their beds, go into the next room, give their reproductions, and the next morning say that they remembered nothing of it" (Jenkins & Dallenbach, 1924, p. 607)! At the time, a semi-asleep state was thought to be similar to hypnosis, so Jenkins and Dallenbach rounded up another student and replicated part of the study, but instead of having the student sleep for varying amounts of time, they had the student learn and recall the lists at different retention intervals while hypnotized (during both learning and recall). They found recall to be virtually perfect for all of the intervals, an early hint at what later came to be called *state-dependent learning* by cognitive psychologists (and similar to Research Example 20, regarding noisy or quiet study environments and exam performance).

Self Test

8.2

1. A maze learning study with a 2 (type of maze: alley maze or elevated maze) × 2 (type of rat: wild or bred in the lab) factorial has these results (DV = number of trials until performance is perfect):

   Alley maze / wild rats = 12 trials
   Alley maze / tame rats = 20 trials
   Elevated maze / wild rats = 20 trials
   Elevated maze / tame rats = 12 trials

   Summarize the results in terms of main effects and interactions.
2. In terms of main effects and interactions, describe the results of the Research Example about studying and taking exams (Research Example 20).

Varieties of Factorial Designs

Like the decision tree in Figure 7.1 for single-factor designs, Figure 8.6 shows the decisions involved in arriving at one of seven factorial designs. You'll recognize that four of the designs mirror those in Figure 7.1, but the other designs are unique to factorials. First, factorial designs can be completely between-subjects, meaning all the independent variables are between-subjects factors. Also, factorial designs can be completely within-subjects, in which all the independent variables are within-subjects factors. When a mixture of between- and within-subjects factors exist in the same experiment, the design is called a **mixed factorial design.** In a mixed design, at least one variable must be tested between subjects, and at least one must be tested within subjects. Second, some between-subjects factorials include both a subject variable and a manipulated independent variable. Because these designs can yield an interaction between the type of person (P) in the study and the situation or environment (E) created in the study, they can be called **P × E factorial designs** ("P by E"), or Person by Environment designs, with Person defined as some subject variable and Environment defined broadly to include any manipulated independent variable. A further distinction can be made

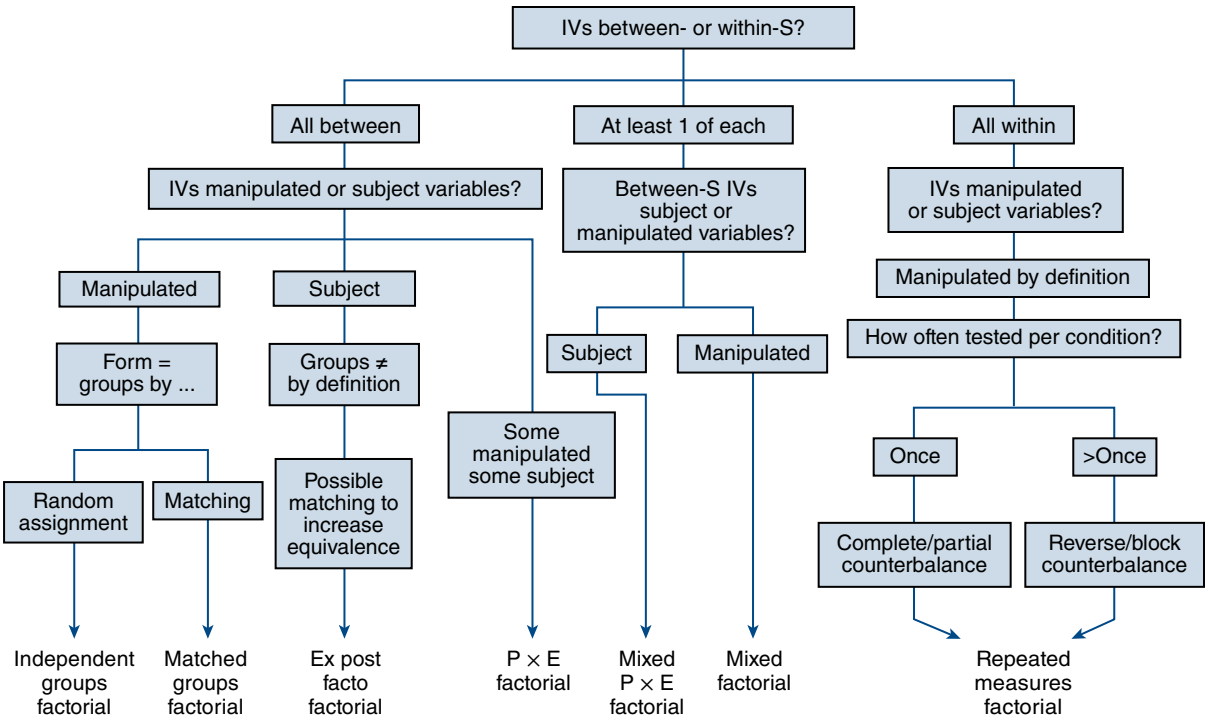


FIGURE 8.6
A decision tree for factorial designs.

within P × E designs, depending on whether the variables are between-subjects or within-subjects factors. In most cases the P variable is a between-subjects factor because it is a subject variable. However, it could be a within-subjects factor if the participants are tested over time, as in a developmental design. The E variable can also be manipulated as either a between- or within-subjects factor. If the P factor is between-subjects, and the E factor is within-subjects, then the P × E is a **mixed P × E factorial**. Let's examine mixed factorials and then P × E factorials in more detail.

Mixed Factorial Designs

In Chapter 6, you learned that when independent variables are between-subjects factors, creating equivalent groups can be a problem, and procedures like random assignment and matching are used to solve the problem. Similarly, when independent variables are within-subjects variables, a difficulty arises because of potential order effects, and counterbalancing is the normal solution. Thus, in a mixed design, the researcher usually gets to deal with both the problems of equivalent groups *and* the problems of order effects. Not always, though—there is one variety of mixed design where counterbalancing is not used because order effects themselves are the outcome of interest. For example, in learning and memory research, "trials" is frequently encountered as a within-subjects factor. Counterbalancing makes no sense in this case because one purpose of the study is to show regular changes from trial to trial. The following two Research Examples show two types of mixed designs, one requiring normal counterbalancing and one in which trials is the repeated measure.

Research Example 21—A Mixed Factorial with Counterbalancing

Terror Management Theory (TMT; Greenberg, Pyszcznski, & Solomon, 1986) is based on the idea that, as a species, we humans have the unique ability to know that our lives will, without a doubt, someday end. This awareness, according to the theory, scares the heck out of us, leading us to develop various coping mechanisms. Concern with death makes it easy for us to believe in some form of afterlife, for example. Research on TMT has focused on seeing what happens if people are reminded of their future death, and these reminders have had a number of interesting effects. For example, Kasser and Sheldon (2000) found that subjects asked to write essays about their future demise developed feelings of insecurity that led them to predict (hope for?) higher estimates of their future financial worth than subjects writing essays about their music preferences.

Cohen, Solomon, Maxfield, Pyszcznski, and Greenberg (2004) examined the relationship between TMT and leadership style. They made the interesting prediction that reminders of death would enhance the appeal of charismatic leaders, leaders whose vision can help overcome feelings of personal insecurity. They designed a 2 × 3 mixed factorial to test the idea. The between-subjects factor, as is typical in TMT studies, randomly assigned participants to one of two groups. One (mortality salient) wrote essays about "the emotions that the thought of your own death arouses in you" (p. 848), while the other (exam salient) wrote essays about the emotions aroused by thoughts of a forthcoming important exam. The within-subjects factor was leadership style. Participants evaluated hypothetical candidates for governor who were described as charismatic (the description emphasized being visionary, creative, and willing to take risks), task-oriented (the description emphasized setting realistic goals and developing clear plans), or relationship-oriented (the description emphasized being friendly and respectful of citizens). You can see why leadership style was tested as a repeated measure—Cohen et al., wanted each participant to evaluate each candidate. With just three levels of the within-subject factor, complete counterbalancing was feasible and was implemented (only six different orders of the three leader descriptions were needed).

The outcome was a significant main effect and an interaction. Although we have seen that interactions sometimes qualify main effects, in this case both outcomes told an important part of the story. The main effect was that task-oriented leaders were generally favored over the other leadership styles, a finding that is fairly typical in research on leadership. The interaction was in line with Cohen et al.'s (2004) TMT prediction. The task-oriented leader was rated highly regardless of mortality or exam salience (middle bars in Figure 8.7); however, reminding participants of their mortality led them to increase their evaluations of charismatic leaders (bars on the left) and decrease their evaluations of relationship-oriented leaders (bars on the right). Apparently, anxiety about the future can lead people to value leaders who promise a secure and optimistic future, and be skeptical about leaders who focus on interpersonal communication.

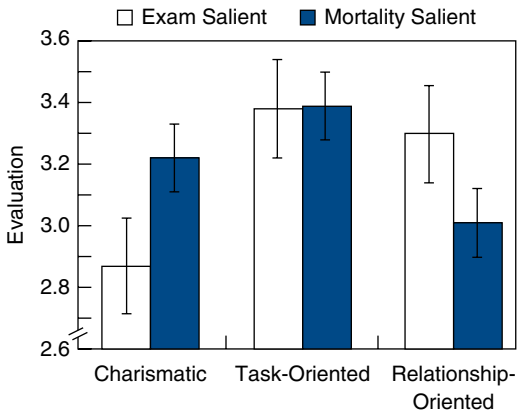


FIGURE 8.7
Mean evaluations of candidates who were described as charismatic, task-oriented, or relationship-oriented (from Cohen, Solomon, Maxfield, Pyszczynski, & Greenberg, 2004)

Research Example 22—A Mixed Factorial without Counterbalancing

In some mixed designs, the within-subjects factor examines changes in behavior with the passage of time—a trials effect—so counterbalancing does not come into play. This was the case in a fascinating study by Crum and Langer (2007); it showed that those doing physical labor for a living might achieve positive health benefits if they did nothing more than redefine their working activities as "exercise." It is no surprise that research shows a relationship between exercise and good health. What you might be surprised to learn is that improvements in health can occur if people merely *think* that what they do in the normal course of a day could be called "exercise."

Crum and Langer (2007) created a 2 × 2 mixed factorial design to examine the extent to which beliefs can affect health. The subjects in the study were 84 women who worked housekeeping jobs in hotels. They work hard, and most of what they do clearly involves exercise (climbing stairs, pushing a heavy cart, lifting mattresses, etc.); Crum and Langer estimated they easily exceed the Surgeon General's recommendation to get 30 minutes of exercise per day to enhance health. Housekeepers don't usually think of their work as exercise, however; they just think of it as work. To alter that perception, Crum and Langer randomly assigned some housekeepers ($n = 44$) to an "informed" group in which they were explicitly told that the work they did exceeded the Surgeon General's recommendations for daily exercise. They were also given specific details about how many calories were burned when doing their tasks (e.g., 15 minutes of vacuuming burns 50 calories). The remaining housekeepers ($n = 40$) were not told anything about their work being considered healthful exercise.[1] The within-subjects factor was time; measures were taken at the start of the study and again four weeks later. There were several dependent variables, both self-report measures (e.g., self-reported levels of exercise when not working) and physiological measures (e.g., blood pressure, body mass index).

Think about this for a minute. Over the course of four weeks, two groups of women did the same amount of physical labor (this was verified independently); the only difference between them was *how they thought about what they were doing*. It is hard to believe that just calling your work "exercise" can have any beneficial physical effect, but that is exactly what happened. After four weeks of thinking their work could also be considered exercise (and therefore healthy), the informed housekeepers showed small but significant decreases in weight, body mass index, body-fat percentage, waist-to-hip ratio, and systolic blood pressure (but not diastolic blood pressure), compared to those in the control group. Mind affects body. Figure 8.8 shows these changes (note the use of error bars in line graphs). In terms of the factorial language we have been using, five of the six graphs (diastolic blood pressure being the exception), showed an interaction between the group (informed or not) and the passage of time. Although it appears there are also main effects for the group factor, at least in the first four graphs, none of these effects were statistically significant.

It might have occurred to you that because housekeepers at a hotel work together, some in the informed group and some in the control group might have talked to each other about the study, thereby muddying the results. This would be an example of *participant crosstalk*, a concept you encountered in Chapter 2. Crum and Langer (2007) thought of this problem and controlled for it in their random assignment procedure. Instead of randomly assigning individuals to one group or another, they randomly assigned *hotels* to the two groups—four hotels in which all the housekeepers were in the informed group, three in which all the housekeepers were in the control group. The hotels in the two groups were similar.

[1] All the subjects in the study knew they were in a study and were told that the purpose was "to find ways to improve the health and happiness of women in a hotel workplace" (p. 167).

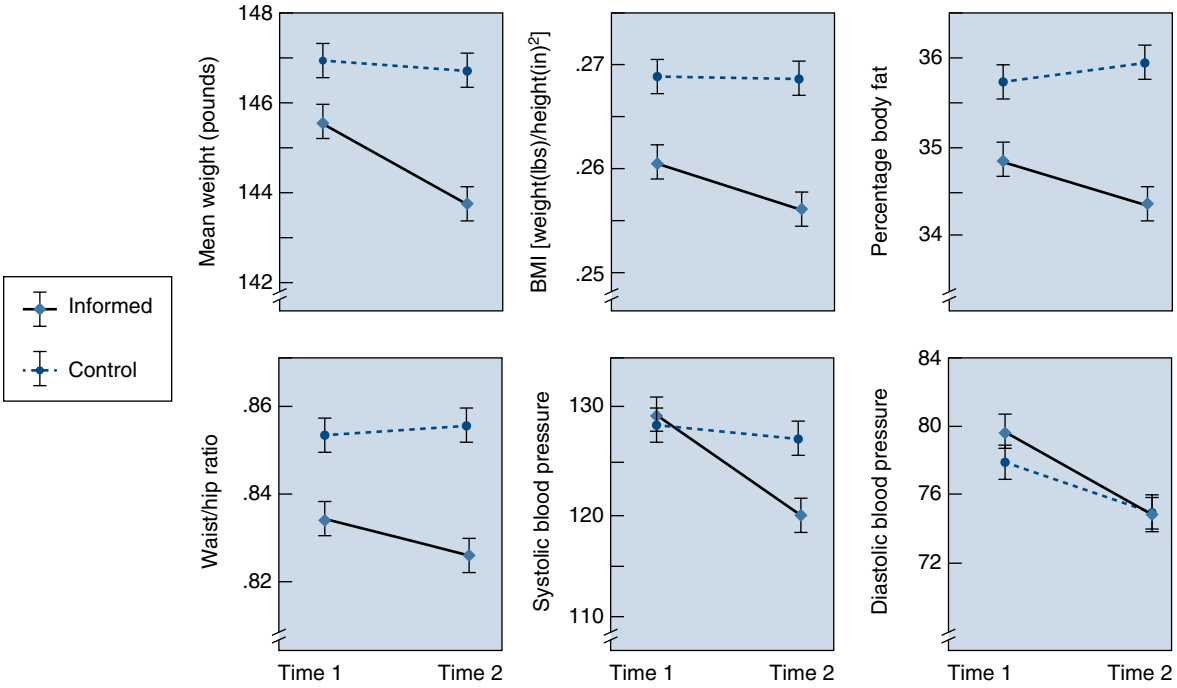


FIGURE 8.8
The health benefits of relabeling work as exercise (from Crum & Langer, 2007).

The final point is an ethical one. After the study was over and it was discovered that the mere relabeling of work as exercise had beneficial health effects for those in the informed group, Crum and Langer (2007) arranged it so all the housekeepers in the control group were given the same information about how their work could be considered healthy exercise.

Factorials with Subject and Manipulated Variables: P × E Designs

Chapter 5 introduced the concept of a subject variable—an existing attribute of an individual such as age, gender, or some personality characteristic. You also learned to be cautious about drawing conclusions when subject variables are involved. Assuming proper control, causal conclusions can be drawn with manipulated independent variables, but with subject variables such conclusions cannot be drawn. P × E designs include both subject and manipulated variables in the same study. Causal conclusions can be drawn if a significant main effect occurs for the manipulated Environment factor, but they cannot be drawn when a main effect occurs for the subject variable or Person factor, and they also cannot be drawn if an interaction occurs. Despite these limitations, designs including both subject and manipulated variables are popular, in part because they combine the two research traditions identified by Woodworth in his famous "Columbia bible" (see the opening paragraphs of Chapter 5). The correlational tradition is associated with the study of individual differences, and the subject variable or *P* factor in the P × E design looks specifically at these differences. A significant main effect for this factor shows two different *types* of individuals perform differently on whatever behavior is being measured as the dependent variable. The experimental tradition, on the other hand, is concerned with identifying

general laws of behavior that apply to some degree to everyone, regardless of individual differences. Hence, finding a significant main effect for the manipulated or *E* factor in a P × E design indicates the situational factor is powerful enough to influence the behavior of many kinds of persons.

Consider a hypothetical example that compares introverts and extroverts (the *P* variable) and asks participants to solve problems in either a small, crowded room or a large, uncrowded room (the *E* variable). Suppose you get results like this (DV = number of problems solved):

		E factor: Small room	E factor: Large room	**Row means**
P factor	Introverts	18	18	**18**
	Extroverts	12	12	**12**
	Column means	**15**	**15**	

In this case, there would be a main effect for personality type, no main effect for environment, and no interaction. Introverts clearly outperformed extroverts (18 > 12) regardless of crowding. The researcher would have discovered an important way in which individuals differ, and the differences extend to more than one kind of environment (i.e., both small and large rooms).

A very different conclusion would be drawn from this outcome:

		E factor: Small room	E factor: Large room	**Row means**
P factor	Introverts	12	18	**15**
	Extroverts	12	18	**15**
	Column means	**12**	**18**	

This yields a main effect for the environmental factor, no main effect for personality type, and no interaction. Here the environment (room size) produced the powerful effect (18 > 12), and this effect extended beyond a single type of individual; regardless of personality type, introverted or extroverted, performance deteriorated under crowded conditions. Thus, finding a significant main effect for the P factor indicates that powerful personality differences occur, while finding a significant main effect for the *E* factor shows the power of some environmental influence to go beyond just one type of person. Of course, another result could be two main effects, indicating that each factor is important.

The most interesting outcome of a P × E design, however, is an interaction. When this occurs, it shows that for one type of individual, changes in the environment have one kind of effect, while for another type of individual, the same environmental changes have a different effect. Staying with the introvert/extrovert example, suppose this happened:

		E factor: Small room	E factor: Large room	Row means
P factor	Introverts	18	12	**15**
	Extroverts	12	18	**15**
	Column means	**15**	**15**	

In this case, neither main effect would be significant, but an interaction clearly occurred. One effect happened for introverts, but something different occurred for extroverts. Specifically, introverts performed much better in the small than in the large room, while extroverts did much better in the large room than in the small one.

Factorial designs that include both subject variables and manipulated variables are popular in educational research and in research on the effectiveness of psychotherapy (Smith & Sechrest, 1991). In both areas, the importance of finding significant interactions is indicated by the fact that such designs are sometimes called **ATI designs**, or "Aptitude-Treatment Interaction designs." As you might guess, the "aptitude" refers to the subject (person) variable and the "treatment" refers to the manipulated, environmental variable. An example from psychotherapy research is a study by Abramovitz, Abramovitz, Roback, and Jackson (1974). Their *P* variable was locus of control. Those with an external locus of control generally believe external events exert control over their lives, while individuals with an internal locus believe what happens to them is a consequence of their own decisions and actions. In the study, externals did well in therapy that was more directive in providing guidance for them, but they did poorly in nondirective therapy, which places more responsibility for progress on the client. For internals, the opposite was true: They did better in the nondirective therapy and not too well in directive therapy.

ATIs in educational research usually occur when the aptitude or person factor is a learning style variable and the treatment or environmental factor is some aspect of instruction. For example, Figure 8.9 shows the outcome of educational research reported by Valerie Shute, a leading authority on ATI designs (Shute, 1994). The study compared two educational strategies for teaching basic principles of electricity: rule induction and rule application. Participants were

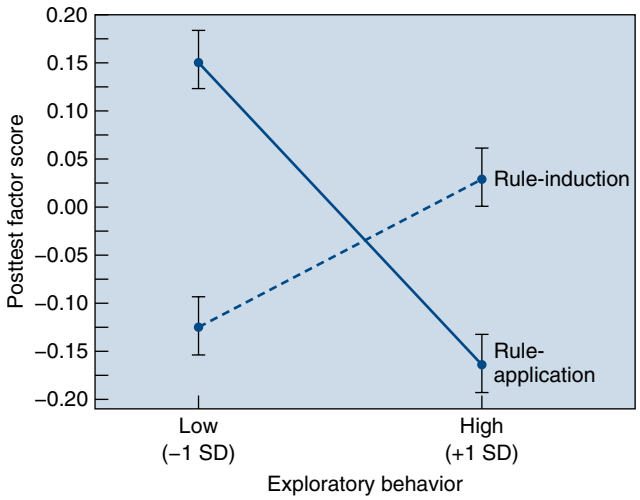


FIGURE 8.9
An ATI interaction between level of exploratory behavior and type of learning environment (from Shute, 1994).

randomly assigned to one strategy or the other. The subject variable was whether learners scored high or low on a measure of "exploratory" behavior. The graph shows those scoring high on exploratory behavior performed better in a rule induction setting, where they were asked to do more work on their own, whereas those scoring low on exploratory behavior performed better in a rule application environment, where the educational procedures were more structured for them.

P × E factorial designs are also popular in research in personality psychology, abnormal psychology, developmental psychology, and any research area interested in gender differences. In personality research, the subject variable or *P* factor will involve comparing personality types; in abnormal psychology, the *P* factor often will be groups of people with different types of mental disorders, and in cross-sectional studies in developmental psychology, the *P* factor will be age. Gender cuts across all of psychology's subdisciplines. The following experiment uses gender as the subject factor and illustrates an unfortunate effect of stereotyping.

Research Example 23—A Factorial Design with a P × E Interaction

Stereotypes are oversimplified and biased beliefs about identifiable groups. They assume that all members of a particular group share particular traits. These traits are usually negative. Stereotypes are dangerous because they result in people being judged with reference to the group they belong to rather than as individuals. The term *stereotype threat* refers to any situation that reminds people of a stereotype. A stereotype that interested Inzlicht and Ben-Zeev (2000) concerns math and the bias that women are not as talented mathematically as men. Their somewhat distressing study shows that stereotypes about women not being suited for math can affect their actual math performance if they are placed in a situation that reminds them of the bias. Experiment 2 of their study was a 2 × 2 P × E factorial. The first factor, the subject variable, was gender; participants were male and female college students at Brown University. The manipulated or environmental factor was the composition of a three-person group given the task of completing a series of math problems. In the "same-sex" condition, all three students taking the test together were either males or females. In the "minority" condition, either women or men were in the minority; that is, the groups included either two men and one woman or two women and one man. The three-person groups had 20 minutes to solve the math problems. They were informed that, after the session was over, their scores would be made public. Figure 8.10 shows the rather startling interaction.

Notice that, for men, performance was unaffected by who was taking the test with them; they did about the same, regardless of whether they took the test with two other men or with two women. It was a different story for women, however. They did quite well when they were in a group of other women (slightly better than the men, in fact), but when they took the test with two other men (the stereotype threat condition), their performance plunged. Keep in mind that the groups didn't even

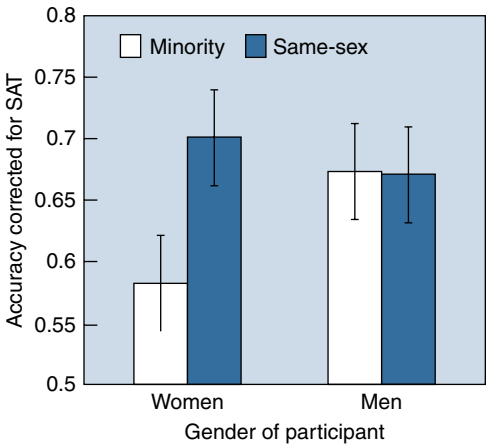


FIGURE 8.10
A P × E interaction: Gender and group composition when solving math problems (adapted from Inzlicht & Ben-Zeev, 2002).

interact; they were just in the same room, taking the test together. Simply being in the room with other men, in the context of a math test, created a perceived stereotype threat and led to a serious drop in performance for women, even women who were highly intelligent (i.e., good enough for admission to a highly selective university). Inzlicht and Ben-Zeev (2000) concluded that the widely held stereotype of men being better at math was evidently sufficient to disrupt the performance of women when they found themselves in the presence of and outnumbered by men. Although they correctly recognized that their study did not directly address the issue of single-sex math classes, Inzlicht and Ben-Zeev argued that "females may in fact benefit from being placed in single-sex math classrooms" (p. 370). On an encouraging note, a more recent study by Johns, Schmader, and Martens (2005) showed that educating women about these stereotype threats, or describing the task as "problem solving" rather than a "math test," substantially reduced the problem.

From the standpoint of the concepts you've been learning about in your methods course, one other point about the Inzlicht and Ben-Zeev (2000) study is worth noting. Remember the concept of *falsification* (Chapters 1, 3, and 5), the process of ruling out alternative hypotheses? The study we just described was actually Experiment 2 of a pair of studies. In Experiment 1, Inzlicht and Ben-Zeev found the drop in women's performance happened when a math test was used, but it did *not* happen with a test of verbal abilities. They contrasted two hypotheses, a "stereotype threat" hypothesis and a "tokenism" hypothesis. The tokenism hypothesis proposes that a person included in a group, but in the minority, perceives herself or himself as a mere token, placed there to give the appearance of inclusiveness. The tokenism hypothesis predicts a decline in female performance *regardless* of the type of test given. Yet the performance decline did not occur with the verbal test, leading Inzlicht and Ben-Zeev to argue that the tokenism hypothesis could be ruled out (falsified), in favor of the alternative, which proposed that performance would decline only in a situation that activated a specific stereotype (i.e., the idea that men and boys are better than women and girls at math).

In this stereotype threat P × E study, the *E* factor (group composition) was tested as a between-subjects factor. If this factor is tested within-subjects, then a P × E design meets the criterion for a mixed design, and can be called a *mixed P × E factorial*. Such is the case in Research Example 24, which includes a rather unsettling conclusion about older drivers.

Research Example 24—A Mixed P × E Factorial with Two Main Effects

Research on the divided attention that results from cell phone use and driving, done by Strayer and his colleagues (Strayer & Johnston, 2001), was mentioned in Chapter 3 as an example of applied research, on attention. Another study by this research team (Strayer & Drews, 2004) compared young and old drivers, a subject variable, and also included two types of tasks: driving while using a cell phone, and driving without using one. This second factor was tested within-subjects—both young and old drivers completed both types of tasks. Hence, the design is a mixed P × E factorial—mixed because it includes both between- (driver age) and within-subjects (cell phone use) factors, and P × E because it includes both a subject (driver age) variable and a manipulated (cell phone use) variable.

Strayer and Drews (2004) operationally defined their subject variable this way: The 20 younger drivers ranged in age from 18 to 25, while the 20 older drivers were between 65 and 74 years old. All the participants were healthy and had normal vision. In a desire to create a procedure with some degree of *ecological validity*, the researchers used a state-of-the-art driving simulator and a "car-following paradigm" (p. 641), in which drivers followed a pace car while other cars passed them periodically. Subjects had to maintain proper distance from the pace car, which would hit the brakes frequently (32 times in a 10-minute trial). The dependent variables were driving speed, distance from the pace car, and reaction time (hitting the brakes when the pace car did). There were four 10-minute trials, two with subjects simply driving ("single-task") and two with the subjects driving while carrying on a cell phone conversation (hands-free phone) with an experimenter ("dual-task"). The cell phone conversations were on topics known from a pre-testing

survey to be of interest to subjects. Because the cell phone factor was a repeated measure, the four trials were counterbalanced "with the constraint that both single- and dual-task conditions were performed in the first half of the experiment and both. . .were performed in the last half of the experiment" (p. 642). You might recognize this as a form of *block randomization* (Chapter 6).

Given the variety of dependent measures, there were several results, but they fell into a general pattern that is best illustrated by the reaction time data. As the factorial matrix here shows, there were main effects for both factors—age and driving condition.

	Driving condition: Single-task	Driving condition: Dual-task	**Row means**
Young drivers	780	912	**846**
Old drivers	912	1086	**999**
Column means	**846**	**999**	

Note: DV = reaction time in ms

Thus, younger drivers (846 ms, or 0.846 seconds) had quicker reactions overall than older drivers (999 ms), and those driving undistracted in the single-task condition (also 846 ms) were quicker overall than those driving while on the cell phone (also 999 ms). There was no interaction. But a look at the cell means yields another result, one with interesting implications: The reaction time for older drivers in the single-task condition is identical to the reaction time for younger drivers in the dual-task condition (912 ms). Reaction time for the young people who were using a cell phone was the *same* as for the old folks not using the phone (!). The outcome is sobering for older drivers (who still think they have it), while at the same time, perhaps of concern to younger drivers. When this study was discussed in class, one creative 20-year-old student asked, "Does this mean that if I talk on my cell while driving, I'll be just like an old person?"

One final point about P × E designs: The label pays homage to the work of Kurt Lewin (1890–1947), a pioneer in social and child psychology. The central theme guiding Lewin's work was that a full understanding of behavior required studying both the person's individual characteristics and the environment in which the person operated. He expressed this idea in terms of a famous formula, $B = f(P, E)$—*B*ehavior is a joint *f*unction of the *P*erson and the *E*nvironment (Goodwin, 2012). P × E factorial designs, named for Lewin's formula, are perfectly suited for discovering the kinds of interactive relationships Lewin believed characterized human behavior.[2]

Recruiting Participants for Factorial Designs

It should be evident from the definitions of factorial design types that the number of subjects needed to complete a study could vary considerably. If you need 5 participants to fill one of the cells in the 2 × 2 factorial, for example, the total number of people to be recruited for the study as a whole could be 5, 10 or 20. Figure 8.11 shows you why. In Figure 8.11a, both variables are tested between subjects, and 5 participants are needed per cell, for a total of 20. In Figure 8.11b, both variables are tested within subjects, making the design a repeated-measures factorial. The same 5 individuals will

[2] One unfortunate implication of Lewin's choice of the label *P* is that a P × E design implies that only human participants are being used. Yet it is quite common for such a design to be used with animal subjects (a study in which the subject variable is the species of the primates being tested, for instance).

(a) For a 2 × 2 design with 4 different groups and 5 participants per cell—20 subjects needed

S1	S11
S2	S12
S3	S13
S4	S14
S5	S15
S6	S16
S7	S17
S8	S18
S9	S19
S10	S20

(b) For a 2 × 2 repeated-measures design with 5 participants per cell—5 subjects needed

S1	S1
S2	S2
S3	S3
S4	S4
S5	S5
S1	S1
S2	S2
S3	S3
S4	S4
S5	S5

(c) For a 2 × 2 mixed design with 5 participants per cell—10 subjects needed

S1	S1
S2	S2
S3	S3
S4	S4
S5	S5
S6	S6
S7	S7
S8	S8
S9	S9
S10	S10

FIGURE 8.11
Participant requirements in factorial designs.

contribute data to each of the four cells. In a mixed design, Figure 8.11c, one of the variables is tested between subjects and the other is tested within subjects. Thus, 5 participants will participate in two cells and 5 will participate in the other two cells, for a total of 10 participants.[3]

Knowing how many participants to recruit for an experiment leads naturally to the question of how to treat the people who arrive at your experiment. Box 8.2 provides a hands-on, practical guide to being an ethically competent researcher.

BOX 8.2 ETHICS—On Being a Competent and Ethical Researcher

You learned about the APA code of ethics in Chapter 2, and you have encountered Ethics boxes in each chapter since then. Although you should have a pretty good sense of the ethical requirements of a study (consent, confidentiality, debriefing, etc.), you might not be sure how to put this into practice. Hence, this might be a good time to give you a list of practical tips for being an ethically responsible experimenter.

- Get to your session early enough to have all of the materials organized and ready to go when your participants arrive.
- Always treat the people who volunteer for your study with the same courtesy and respect you would hope to receive if the roles were reversed. Greet them when they show up at the lab and thank them for signing up and coming to the session. They might be apprehensive about what will happen to them in a psychology experiment, so your first task is to put them at ease, at the same time maintaining your professional role as the person in charge of the session. Always remember that they are doing you a favor—the reverse is not true. Smile often.
- Start the session with the informed consent form. Don't convey the attitude that this is a time-consuming technicality that must be completed before the important part starts. Instead, make it clear you want your participants to have a good idea of what they are being asked to do. If they don't ask questions while reading the consent form, be sure to ask them if they have any when they finish reading. Make sure there are two copies of the signed consent form, one for them to take and one for your records.
- Prepare a written *protocol* (see Chapter 6) ahead of time. This is a detailed sequence of steps you must complete in order to run the session successfully from start to finish. It helps ensure each subject has a standardized experience. The protocol might include explicit instructions to be read to the subjects, or it might indicate the point where subjects are given a sheet of paper with written instructions on it.
- Before you test any "real" participants, practice playing the role of experimenter a few times with friends or lab partners. Go through the whole experimental procedure. Think of it as a dress rehearsal and an opportunity to iron out any problems with the procedure.
- Be alert to signs of distress in your subjects during the session. Depending on the constraints of the procedure, this could mean halting the study and discarding their data, but their welfare is more important than your data. Also, you are not a professional counselor; if they seem disturbed by their participation, gently refer them to your course instructor or the school's counseling center.
- Prepare the debriefing carefully. As a student experimenter, you probably won't be running studies involving elaborate deception or producing high levels of stress, but you will be responsible for making this an educational experience for your participants. Hence, you should work hard on a simplified description of what the study hopes to discover, and you should give participants the chance to suggest improvements in the procedure or ideas for the next study. So don't rush the debriefing or give a cursory description that implies you hope they will just leave. And if they seem to want to leave without any debriefing (some will), try not to let them. Debriefing is an important part of your responsibility as a researcher and an educator. (Of course, if they say, "I thought you said we could leave any time," there's not much you can do!)

[3] These small sample sizes are used merely to illustrate the subject needs for the types of factorial designs. In actual practice, sample sizes are typically much larger, determined either through a power analysis (Chapter 4) or with reference to standard practice in some research area. In a recent article that made several proposals for avoiding false positives (Type I errors) in research, Simmons, Nelson, and Simonsohn (2011) recommended that studies "collect a minimum of 20 observations per cell or else provide a compelling cost-of-data-collection justification" (p. 1363).

- Before they go, remind participants that the information on the consent form includes names of people to contact about the study if questions occur to them later. Give them a rough idea of when the study will be completed and when they can expect to hear about the overall results (if they indicate they would like to receive this information). To avoid *participant crosstalk* (Chapter 2), ask them not to discuss the experiment with others who might be participants. Participant crosstalk can be a serious problem, especially at small schools (refer to Box 6.3 in Chapter 6, for more on the responsibilities of research subjects). If you are good to your participants throughout the session, however, you increase the chances of their cooperation in this regard. Also, remember from Chapter 2 that if you have special reasons to be concerned about crosstalk (e.g., substantial deception in a study), the ethics code allows you to make the debriefing more cursory, as long as you give them the opportunity to receive complete results once the study is completed.
- As they are leaving, be sure to thank them for their time and effort, and be sure you are smiling as they go out the door. Remember that some of the students you test will be undecided about a major and perhaps thinking about psychology. Their participation in your study could enhance their interest.

Analyzing Data from Factorial Designs

We have already seen that multilevel, single-factor designs using interval or ratio data are analyzed using an analysis of variance (ANOVA) procedure. ANOVAs are also the analysis of choice for factorial designs. When doing a one-way ANOVA, just one *F* ratio is calculated. Then subsequent testing may be done if the *F* is significant. For a factorial design, however, more than one *F* ratio will be calculated. Specifically, there will be an *F* for each possible main effect and for each possible interaction. For example, in the 2 × 2 design investigating the effects of imagery training and presentation rate on memory, an *F* ratio will be calculated to examine the possibility of a main effect for type of training, another for the main effect of presentation rate, and a third for the potential interaction between the two. In an A × B × C factorial, *seven F* ratios will be calculated: three for each of the main effects of A, B, and C; three more for the two-way interaction effects of A × B, B × C, and A × C; plus one for the three-way interaction, A × B × C.

As you recall from Chapter 7, the type of design dictates whether the one-way ANOVA will be an ANOVA for independent groups or a repeated measures ANOVA. In the same way, the design also determines if a factorial ANOVA will be one of these two types, or a third type: A mixed ANOVA is called for when a mixed factorial design is used. Also, as was the case for one-way ANOVAs, subsequent (post hoc) testing may occur with factorial ANOVAs. For example, in a 2 × 3 ANOVA, a significant main effect for the factor with three levels would trigger a subsequent analysis (e.g., Tukey's HSD) that compared the overall performance of levels 1 and 2, 1 and 3, and 2 and 3. Following a significant interaction, one common procedure is to complete a **simple effects analysis**. This involves comparing each of the levels of one factor with each level of the other factor. A concrete example will make this clear. Refer to the point in the chapter where we introduced interactions by discussing a 2 × 2 factorial with type of course (lab versus lecture) and type of student (science versus humanities major) as the factors. As you recall, no main effects occurred (row and column means were all 75). A simple effects analysis would make these comparisons:

1. For science majors, compare lab emphasis (mean of 80) with lecture emphasis (70)
2. For humanities majors, compare lab emphasis (70) with lecture emphasis (80)
3. For the lab emphasis, compare science (80) with humanities majors (70)
4. For the lecture emphasis, compare science (70) with humanities majors (80)

For details on how to complete a simple effects analysis, consult any good statistics text (e.g., Witte & Witte, 2010).

For information on how to create and ANOVA source table for a 2 × 2 factorial ANOVA for independent groups, consult the Student Statistics Guide on the Student Companion Site. A good statistics text will explain how to create source tables for other forms of ANOVA. In addition, see the guide to learn how to perform various factorial ANOVAs using SPSS.

SELF TEST

8.3

1. What is the defining feature of a mixed design? In a 3 × 3 mixed design with 20 subjects in the first cell, how many subjects are needed to complete the study?
2. Distinguish a P × E design from an ATI design.
3. If you need a total of 25 participants in a 4 × 4 factorial study and there are 25 participants in one of the cells, what kind of design is this?

Before closing this chapter, here is one final point about factorial designs and the analysis of variance. You've been looking at many factorial matrices in this chapter. They might vaguely remind you of aerial views of farms in Kansas. If so, it's no accident, as you can discover by reading Box 8.3, which tells you a bit about Sir Ronald Fisher, who invented the analysis of variance.

BOX 8.3 ORIGINS—Factorials Down on the Farm

Imagine you're in a small plane flying over Kansas. Looking out the window, you see mile after mile of farms, their fields laid out in blocks. The pattern might remind you of the factorial matrices you've just encountered in this chapter. This is probably a coincidence, but factorial designs and the ANOVA procedures for analyzing them were first developed in the context of agricultural research by Sir Ronald Fisher. The empirical question was, "What are the best possible conditions or combinations of conditions for raising crop X?"

Ronald Aylmer Fisher (1890–1962) was one of Great Britain's best-known statisticians, equal in rank to the great Karl Pearson, who invented the correlation measure we now call Pearson's *r* (next chapter). Fisher created statistical procedures useful in testing predictions about genetics, but he is perhaps best known among research psychologists for creating the ANOVA, which yielded *F* ratios that allowed decisions about the null hypothesis in experimental agricultural research. You can easily guess what the *F* represents.

For about 15 years beginning in 1920, Fisher worked at an experimental agricultural station at Rothamsted, England. While there, he was involved in research investigating the effects on crop yield of such variables as fertilizer type, rainfall level, planting sequence, and genetic strain of various crops. He published articles with titles like "Studies in Crop Variation: VI. Experiments on the Response of the Potato to Potash and Nitrogen" (Kendall, 1970, p. 447). In the process, he invented ANOVA as a way of analyzing the data. He especially emphasized the importance of using factorial designs, "for with separate [single-factor] experiments we should obtain no light whatever on the possible *interactions* of the different ingredients" (Fisher, 1935/1951, p. 95, italics added). In the real world of agriculture, crop yields resulted from complex combinations of factors, and studying one factor at a time wouldn't allow a thorough evaluation of those interactive effects. As you have seen in this chapter, the interaction is often the most intriguing result in a factorial study.

A simple 2 × 2 design for one of Fisher's experiments, with each block representing how a small square of land was treated, might look like Figure 8.12. As with any factorial, this design allows one to evaluate main effects (of fertilizer

and type of wheat in this case), as well as the interaction of the two factors. In the example in Figure 8.12, if we assume the shaded field produces significantly more wheat than the other three (which equal each other), then we would say an interaction clearly occurred: The fertilizer was effective, but for only one specific strain of wheat.

Fisher first published his work on ANOVA in book form in 1925 (a year after Jenkins and Dallenbach published their classic sleep and memory study), as part of a larger text on statistics (Fisher, 1925). His most famous work on ANOVA, which combined a discussion of statistics and research methodology, appeared 10 years later as *The Design of Experiments* (Fisher, 1935/1951). ANOVA techniques and factorial designs were slow to catch on in the United States, but by the early 1950s, they had become institutionalized as a dominant statistical tool for experimental psychologists (Rucci & Tweney, 1980).

	Experimental fertilizer	No experimental fertilizer
Wheat: genetic strain I	Wheat field A	Wheat field B
Wheat: genetic strain II	Wheat field C	Wheat field D

FIGURE 8.12
An agricultural interaction.

This completes our two-chapter sequence about experimental design. The material (along with Chapters 5 and 6) is sure to require more than one reading and a fair amount of practice with designs before you'll feel confident about your ability to use experimental psychologist language fluently and to create a methodologically sound experiment that is a good test of your hypothesis. Next up is a closer look at a research tradition in which the emphasis is not on examining differences but degrees of association between measured variables.

CHAPTER SUMMARY

Essentials of Factorial Designs

Factorial designs examine the effects of more than one independent variable. Factorial designs are identified with a notation system that identifies the number of independent variables, the number of levels of each independent variable, and the total number of conditions in the study. For example, a 2 × 3 ("2 by 3") factorial design has two independent variables, the first with two levels and the second with three levels, and six different conditions (2 times 3).

Outcomes—Main Effects and Interactions

The overall influence of an independent variable in a factorial study is called a main effect. There are two possible main effects in a 2 × 3 design, one for the factor with two levels and one for the factor with three levels. The main advantage of a factorial design over studies with a single independent variable is that factorials allow the discovery of interactions between the factors. In an interaction, the influence of one independent variable differs for the levels of the other independent variable. The outcomes of factorial studies can include significant main effects, interactions, both, or neither. When a study yields both main effects and interactions, the interactions should be interpreted first; sometimes an interaction is the important result, while the main effects in the study are irrelevant.

Varieties of Factorial Designs

All of the independent variables in a factorial design can be between-subjects factors or all can be within-subjects factors. Completely between-subjects factorial designs can include independent groups, matched groups, or ex post facto designs. Completely within-subjects factorial designs are also called repeated-measures factorial designs. A mixed factorial design includes at least one factor of each type (between and within). Factorial designs with at least one subject variable and at least one manipulated variable allow for the discovery of Person × Environment (P × E) interactions. When these interactions occur, they show how stimulus situations affect one type of person one way and a second type of person another way. A main effect for the *P* factor (i.e., subject variable) indicates important differences between types of individuals that exist in several environments. A main effect for the *E* factor (i.e., manipulated variable) indicates important environmental influences that exist for several types of persons. In educational research and research on the effectiveness of psychotherapy, these interactions between persons and environments are sometimes called Aptitude-Treatment-Interactions (ATIs). In a mixed P × E design, the *E* factor is a within-subjects variable.

CHAPTER REVIEW QUESTIONS

1. For a factorial design, distinguish between "levels" and "conditions."
2. What is meant by a main effect? In terms of the contents of a factorial matrix, how does one go about determining if a main effect has occurred?
3. Use the "closing time" study by Gladue and Delaney (1990) to show that an experiment can result in two important outcomes— two main effects.
4. Use the Grant et al. (1998) experiment (studying in noisy or quiet environments) to show that important results can occur in a study, even if no main effects occur.
5. In a study with both main effects and an interaction, explain why the interaction must be interpreted first and how the statistically significant main effects might have little meaning for the overall outcome of the study. Use the caffeine study to illustrate.
6. Distinguish between a mixed factorial design and a P × E design. How can a design be both a mixed design and a P × E design?
7. Use the introvert/extrovert and room size example to show how P × E designs can discover important ways in which (a) individuals differ, and (b) situations can be more powerful than individual differences.
8. Mixed factorial designs may or may not involve counterbalancing. Explain.
9. Describe the basic research design and the general outcome of Jenkins and Dallenbach's (1924) famous study on sleep and memory. What interaction might have occurred in their study?
10. What is a simple effects analysis and when are these analyses done?

APPLICATIONS EXERCISES

Exercise 8.1. Identifying Designs

For each of the following descriptions of studies, identify the independent and dependent variables involved, the levels of the independent variable, and the nature of each independent variable (between-subjects or within-subjects; manipulated or subject variables). Also, describe the number of independent variables and levels of each by using the factorial notation system (e.g., 2 × 3), and use Figure 8.6 to identify the design.

1. On the basis of scores on the Jenkins Activity Survey, three groups of subjects are identified: Type A, Type B, and intermediate. An equal number of subjects in each group are given one of two tasks to perform. One of the tasks is to sit quietly in a small room and estimate, in the absence of a clock, when 2 full minutes have elapsed. The second task is to make the same estimate, except that while in the small room, the subject will be playing a hand-held video game.
2. College students in a cognitive mapping study are asked to use a direction finder to point accurately to three unseen locations that vary in distance from the lab. One is a nearby campus location, one is a nearby city, and the third is a distant city. Half of the participants perform the task in a windowless room with a compass indicating the direction of north. The remaining participants perform the task in the same room without a compass.
3. In a study of touch sensitivity, two-point thresholds are measured on 10 skin locations for an equal number of blind and sighted adults. Half of the participants perform the task in the morning and half in the evening.
4. Three groups of preschoolers are put into a study of delay of gratification in which the length of the delay is varied. Children in all three groups complete a puzzle task. One group is told that as payment they can have $1 now or $3 tomorrow. The second group chooses between $1 now and $3 two days from now, and the third group chooses between $1 now and $3 three days from now. For each of the three groups, half of the children solve an easy puzzle and half solve a difficult puzzle. The groups are formed in such a way that the average parents' income is the same for children in each group.
5. In a study of visual illusions and size perception, participants adjust a dial that alters one of two stimuli. The goal is to make the two stimuli appear equal in size, and the size of the error in this judgment is measured on each trial. Each participant completes 40 trials. On half of the trials, the pairs of stimuli are in color; on the other half, they are in black and white. For both the colored and the black-and-white stimuli, half are presented at a distance of 10 feet from the participant and half are presented at 20 feet.
6. In a study of reading comprehension, sixth-grade students read a short story about baseball. The students are divided into two groups based on their knowledge of baseball. Within each group, half of the students are high scorers on a test of verbal IQ, while the remaining students are low scorers.

7. In a study on stereotyping, students are asked to read an essay said to be written by either a psychiatric patient or a mental health professional. Half of the subjects given each essay are told the writer is a male and half are told the writer is a female. Subjects are randomly assigned to the four groups and asked to judge the quality of the essay.

8. In a maze learning study, the performance (number of trials to learn the maze) of wild and lab-reared albino rats is compared. Half of each group of rats is randomly assigned to an alley maze; others learn an elevated maze.

Exercise 8.2. Main Effects and Interactions

For each of the following studies:

a. Identify the independent variables, the levels of each, and the dependent variable.

b. Place the data into the correct cells of a factorial matrix and draw a graph of the results.

c. Determine if main effects and/or interactions exist and give a verbal description of the study's outcome.

For the purposes of the exercise, assume that a difference of *more than 2* between any of the row, column, or cell means is a statistically significant difference.

1. A researcher is interested in the effects of ambiguity and number of bystanders on helping behavior. Participants complete a questionnaire in a room with zero or two other people (i.e., bystanders) who appear to be other subjects but are actors in the study. The experimenter distributes the questionnaire and then goes into the room next door. After 5 minutes, there is a loud crash, possibly caused by the experimenter falling. For half of the participants, the experimenter unambiguously calls out that he has fallen, is hurt, and needs help. For the remaining participants, the situation is more ambiguous; the experimenter says nothing after the apparent fall. In all cases, the actors (bystanders) do not get up to help. The experimenter records how long it takes (in seconds) before a participant offers help. Here are the four conditions and the data:

0 bystanders, ambiguous	24 sec
2 bystanders, ambiguous	38 sec
0 bystanders, unambiguous	14 sec
2 bystanders, unambiguous	14 sec

2. In a maze learning study, a researcher is interested in the effects of reinforcement size and reinforcement delay. Half of the rats in the study are given a 1 cm square block of cheese upon completing the maze; the other half gets a 2 cm square block. Within each reinforcement size group, half of the rats are given the cheese on arrival at the goal box and half waits for the cheese for 15 seconds after their arrival. Hence, there are four groups and the data (dependent variable is number of errors during 10 trials):

small reward, 0 sec delay	17 errors
large reward, 0 sec delay	15 errors
small reward, 15 sec delay	25 errors
large reward, 15 sec delay	23 errors

3. A cognitive psychologist interested in gender and spatial ability decides to examine whether gender differences in a mental rotation task (see Chapter 4 for a reminder of this task) can be influenced by instructions. One set of instructions emphasizes the spatial nature of the task and relates it to working as a carpenter (male-oriented instructions); a second set of instructions emphasizes the problem-solving nature of the task and relates it to working as an interior decorator (female-oriented instructions); the third set of instructions is neutral. An equal number of men and women participate in each instructional condition. Here are the six conditions and the data:

men with male-oriented instructions	26 problems correct
men with female-oriented instructions	23 problems correct
men with normal instructions	26 problems correct
women with male-oriented instructions	18 problems correct
women with female-oriented instructions	24 problems correct
women with normal instructions	18 problems correct

4. A forensic psychologist wishes to determine if prison sentence length can be affected by defendant attractiveness and facial expression. Subjects read a detailed crime description (a felony breaking and entering) and are asked to recommend a sentence for the criminal, who has been arrested and found guilty. A photo of the defendant accompanies the description. Half the time the photo is of a woman made up to look attractive, and half the time the woman is made up to look unattractive. For each type of photo, the woman is smiling, scowling, or showing a neutral expression. Here are the conditions and the data:

attractive, smiling	8 years
attractive, scowling	14 years
attractive, neutral	9 years
unattractive, smiling	12 years
unattractive, scowling	18 years
unattractive, neutral	13 years

Exercise 8.3. Estimating Participant Needs

For each of the following, use the available information to determine how many research subjects are needed to complete the study (*Hint*: One of these is unanswerable without more information):

1. a 3 × 3 mixed factorial; each cell needs 10 participants
2. a 2 × 3 repeated-measures factorial; each cell needs 20 participants
3. a 2 × 4 mixed factorial; each cell needs 8 participants
4. a 2 × 2 × 2 independent groups factorial; each cell needs 5 participants
5. a 2 × 2 matched groups factorial; each cell needs 8 participants
6. a 4 × 4 ex post facto factorial; each cell needs 8 participants

Answers to Self Tests

✓**8.1**

1. (a) 3; (b) 2, 3, and 4; (c) 24
2. A main effect concerns whether a significant difference exists among the levels of an independent variable.
3. A main effect for the type of instruction factor (row means of 20 for imagery and 12 for rote), but no main effect for presentation rate (both column means are 16).

✓**8.2**

1. There are no main effects (row and column means all equal 16), but there is an interaction. Wild rats performed better (fewer trials to learn) in the alley maze, while tame rats performed better in the elevated maze.
2. No overall main effect for whether studying took place in a noisy or a quiet environment; also, no overall main effect for whether recall took place in noisy or quiet environment; there was an interaction—recall was good when study and test conditions matched, and poor when study and test conditions did not match.

✓**8.3**

1. There is at least one between-subjects factor and at least one within-subjects factor; 60.
2. A P × E design has at least one subject factor (P) and one manipulated (E) factor; an ATI design is a type of P × E design in which the "P" factor refers to some kind of ability or aptitude; these ATI designs are frequently seen in educational research.
3. It must be a repeated-measure factorial design.