Methodological Control in Experimental Research

6

PREVIEW & CHAPTER OBJECTIVES

In Chapter 5, you learned the essentials of the experimental method—manipulating an independent variable, controlling extraneous variables, and measuring the dependent variable. In this chapter, we examine two general types of experimental designs, one in which different groups of subjects contribute data to different levels of the independent variable (between-subjects design) and one in which the same subjects contribute data to all levels of the independent variable (within-subjects design). As you are about to learn, each approach has advantages, but each has a problem that must be carefully controlled: the problem of equivalent groups for between-subjects designs and problem of order effects for within-subjects designs. The last third of the chapter addresses the issue of experimenter and participant biases and ways of controlling them. When you finish this chapter, you should be able to:

- Distinguish between-subjects designs from within-subjects designs.
- Understand how random assignment solves the equivalent groups problem in between-subjects designs.
- Understand when matching, followed by random assignment, should be used instead of simple random assignment when attempting to create equivalent groups.
- Understand why counterbalancing is needed to control for order effects in within-subjects designs.
- Distinguish between progressive and carry-over effects in within-subjects designs, and understand why counterbalancing usually works better with the former than with the latter.
- Describe the various forms of counterbalancing for situations in which participants are tested once per condition and more than once per condition.
- Describe the specific types of between- and within-subjects designs that occur in research in developmental psychology, and understand the methodological problems associated with each.
- Describe how experimenter bias can occur and how it can be controlled.
- Describe how participant bias can occur and how it can be controlled.
- Describe the origins of the biasing phenomenon known as the Hawthorne effect.

In his landmark experimental psychology text, just after introducing his now famous distinction between independent and dependent variables, R. S. Woodworth emphasized the importance of *control* in experimental research. As Woodworth (1938) put it, "Whether one or more independent variables are used, it remains essential that all other conditions be constant. Otherwise you cannot connect the effect observed with any definite cause. The psychologist must expect to encounter difficulties in meeting this requirement." (p. 3). Some of these difficulties we have already seen. The general problem of confounding and the specific threats to internal validity, discussed in the previous chapter, are basically problems of controlling extraneous factors. In this chapter, we will describe other aspects of maintaining control: the problem of creating equivalent groups in experiments involving separate groups of subjects, the problem of order effects in experiments in which subjects are tested several times, and problems resulting from biases held by both experimenters and research participants.

Recall that any independent variable must have a minimum of two levels. At the very least, an experiment will compare level A with level B. Those who participate in the study might be placed in level A, level B, or both. If they receive either A *or* B but not both, the design is a **between-subjects design**, so named because the comparison of conditions A and B will be a contrast *between* two groups of individuals. On the other hand, if each participant receives both levels A *and* B, you could say both levels exist *within* each individual participating in the study; hence, this design is called a **within-subjects design** (or, sometimes, a *repeated-measures design*). Let's examine each approach.

Between-Subjects Designs

Between-subjects designs are sometimes used because they must be used. If the independent variable is a subject variable, for instance, there is usually no choice. A study comparing introverts with extroverts requires two different groups of people, some shy, some outgoing; a study on gender differences in children requires a group of girls and a group of boys.

Using a between-subjects design is also unavoidable in studies that use certain types of manipulated independent variables. That is, sometimes when people participate in one level of an independent variable, the experience gained there will make it impossible for them to participate in other levels. This often happens in social psychological research and most research involving deception. Consider an experiment on the effects of the physical attractiveness of a defendant on recommended sentence length by Sigall and Ostrove (1975). They gave college students descriptions of a crime and asked them to recommend a jail sentence for the woman convicted. There were two separate between-subjects, manipulated independent variables. One variable was the type of crime—either a burglary in which "Barbara Helm" broke into a neighbor's apartment and stole $2,200 (a fair amount of money in 1975) or a swindle in which Barbara "ingratiated herself to a middle-aged bachelor and induced him to invest $2,200 in a nonexistent corporation" (p. 412). The other manipulated variable was Barbara's attractiveness. Some participants saw a photo of her in which she was very attractive, others saw a photo of a Barbara made up to be unattractive (the same woman posed for both photos), and a control group did not see any photo. The interesting result was that when the crime was burglary, attractiveness paid. Attractive Barbara got a lighter sentence on average (2.80 years) than unattractive (5.20) or control (5.10) Barbara. However, the opposite happened when the crime was a swindle. Apparently thinking Barbara was using her good looks to commit the crime, participants gave attractive Barbara a

harsher sentence (5.45 years) than they gave the unattractive (4.35) or control (4.35) Barbara. (Notice we are reporting the *means* for each condition here; we will do this when we describe many studies in this textbook.)

Can you see why it was necessary to run this study with between-subjects independent variables? For those participating in the Attractive-Barbara-Swindle condition, for example, the experience would certainly affect them and make it impossible for them to start fresh in, say, the Unattractive-Barbara-Burglary condition. In some studies, participating in one condition makes it impossible for the same person to be in a second condition. Sometimes, it is essential that each condition include uninformed participants.

While the advantage of a between-subjects design is that each subject enters the study fresh, and naïve with respect to the hypotheses to be tested, the prime disadvantage is that large numbers of people may need to be recruited, tested, and debriefed during the course of the experiment. Hence, the researcher invests a great deal of energy in this type of design. The Barbara Helm study used six groups with 20 subjects in each group for a total of 120 people.

Another disadvantage of between-subjects designs is that differences between the conditions might be due to the independent variables, but they might also be due to differences between the individuals in the different groups. Perhaps the subjects in one group are smarter than those in another group. To deal with this potential confound, deliberate steps must be taken to create **equivalent groups**. These groups are equal to each other in every important way except for the levels of the independent variable.

Creating Equivalent Groups

There are two common techniques for creating equivalent groups in a between-subjects experiment. One approach is to use simple random assignment. A second strategy is to use a matching procedure, followed by random assignment.

Random Assignment

First, be sure you understand that *random assignment* and *random selection* are not the same procedures. Random selection is a sampling procedure designed to obtain a *random sample* as described in Chapter 4. It is a process designed to produce a sample of individuals that reflects the broader population, and it is a common strategy used in survey research. Random assignment, in contrast, is a method for placing participants, once already selected for a study, into the different conditions. When **random assignment** is used, every person volunteering for the study has an equal chance of being placed in any of the conditions being formed.

The goal of random assignment is to take individual difference factors that could influence the study and spread them evenly throughout the different groups. For instance, suppose you are comparing two presentation rates in a simple memory study. Further suppose anxious participants don't do as well on your memory task as non-anxious participants, but you as the researcher are unaware of that at the outset of the study (or it just doesn't occur to you). Some subjects are shown a word list at a rate of 2 seconds per word; others at 4 seconds per word. The prediction is that recall will be better with a longer presentation rate or for the 4-second condition. You randomly assign participants to one condition or the other. Here are some hypothetical data that such a study might produce. Each number refers to the number of words recalled out of a list of 30. After each subject number, we placed an *A* or an *R* in parentheses to indicate which participants are anxious (A) and which are relaxed (R). Data for the anxious people are shaded.

Participant	*2-Second Rate*	*Participant*	*4-Second Rate*
S1(R)	16	S9 (R)	23
S2(R)	15	S10 (R)	19
S3(R)	16	S11 (R)	19
S4(R)	18	S12 (R)	20
S5(R)	20	S13 (R)	25
S6(A)	10	S14 (A)	16
S7(A)	12	S15 (A)	14
S8(A)	13	S16 (A)	16
M	**15.00**	***M***	**19.00**
SD	**3.25**	***SD***	**3.70**

If you look carefully at these data, you'll see the three anxious participants in each group did worse than their five relaxed peers. Because the number of anxious participants in each group is equal, however, the dampening effect of anxiety on recall is about the same for both groups. Thus, the main comparison of interest, the difference in presentation rates, is preserved—a mean of 15 words for the 2-second group and 19 for the 4-second group.

Random assignment won't guarantee placing an equal number of anxious participants in each group, but in general, the procedure has the effect of spreading potential confounds evenly among the groups. This is especially true when large numbers of individuals are assigned to each group. In fact, the greater the number of subjects involved, the greater the chance that random assignment will work to create equivalent groups. If groups are equivalent and everything else is adequately controlled, then you are in the enviable position of being able to say your independent variable likely caused differences between your groups.

You might think the process of random assignment would be fairly simple, but the result of such a procedure is that your groups will almost certainly contain different numbers of people. In the worst-case scenario, imagine you are doing a study using 20 participants divided into two groups of 10. You decide to flip a coin as each volunteer arrives: heads, they're in group A; tails, group B. But what if the coin comes up heads all 20 times? Unlikely, but possible.

To complete the assignment of participants to conditions in a way that guarantees an equal number of subjects per group, a researcher can use **blocked random assignment**, a procedure ensuring that each condition of the study has a participant randomly assigned to it before any condition is repeated a second time. Each block contains all of the conditions of the study in a randomized order. This can be done by hand, using a table of random numbers, but researchers typically rely on a simple computer application to generate a sequence of conditions meeting the requirements of block randomization; you can find one at www.randomizer.org that will accomplish both random sampling and random assignment.

One final point about random assignment is that the process is normally associated with laboratory research. That environment allows a high degree of control, so it is not difficult to ensure that each person signing up for the study has an equal chance of being assigned to any of the conditions. Although it is not always feasible, random assignment is also possible in some field research. For example, some universities use random assignment to assign roommates, thereby providing an opportunity to study a number of factors affecting college students. For example,

Shook and Fazio (2008) examined the so-called contact hypothesis, the idea that racial prejudice can be reduced when members of different races are in frequent contact with each other (and other factors, such as equal status, are in play). They found a university where roommates were randomly assigned and designed an experiment to compare two groups of white first-year students—those randomly assigned to another white student and those randomly assigned to an African-American student. Over the course of a fall quarter, the researchers examined whether the close proximity of having a different-race roommate would reduce prejudice. In line with the contact hypothesis, it did.

Matching

When only a small number of subjects are available for your experiment, random assignment can, by chance, fail to create equivalent groups. The following example shows how this might happen. Let's take the same study of the effect of presentation rate on memory, used earlier, and assume the data you just examined reflect an outcome in which random assignment happened to work—that is, there was an exact balance of five relaxed and three anxious people in each group. However, it is *possible* that random assignment could place all six of the anxious participants in *one* of the groups. This is unlikely, but it could occur (just as it's remotely possible for a perfectly fair coin to come up heads 10 times in a row). If it did, this might happen:

Participant	*2-Second Rate*	*Participant*	*4-Second Rate*
S1(R)	15	S9 (R)	23
S2(R)	17	S10 (R)	20
S3(R)	16	S11 (A)	16
S4(R)	18	S12 (A)	14
S5(R)	20	S13 (A)	16
S6(R)	17	S14 (A)	16
S7(R)	18	S15 (A)	14
S8(R)	15	S16 (A)	17
M	**17.00**	***M***	**17.00**
SD	**1.69**	***SD***	**3.07**

This outcome, of course, is totally different from the first example. Instead of concluding that recall was better for a slower presentation rate (as in the earlier example), the researcher in this case could not reject the null hypothesis (that is, the groups are equal: 17 = 17). Participants were randomly assigned, and the researcher's prediction about better recall for a slower presentation rate certainly makes sense. So what went wrong?

Random assignment, in this case, inadvertently created two decidedly nonequivalent groups—one made up entirely of relaxed people and one mostly including anxious folks. A 4-second rate probably does produce better recall, but the true difference was not found in this study because the mean for the 2-second group was inflated by the relatively high scores of the relaxed participants and the 4-second group's mean was suppressed because of anxiety. Another way of saying this is that the failure of random assignment to create equivalent

groups probably led to a Type II error (presentation rate really does affect recall but this study failed to find the effect). To repeat what was mentioned earlier, a critical point is that the chance of random assignment working to create equivalent groups increases as sample size increases. Our example only had 16 participants, so there is a good chance of creating non-equivalent groups even using random assignment.

Note the exact same outcome just described could also occur if you failed to make any effort to create equivalent groups. For example, suppose you tested people as they signed up for your study, starting with the 2-second/item condition and then finishing with the 4-second/item condition. It is conceivable that anxious subjects would be slower to sign up than relaxed subjects, resulting in the second group being composed mostly of anxious subjects.

A second general strategy for deliberately trying to create equivalent groups is to use a matching procedure. In **matching**, participants are grouped together on some *subject variable* such as their characteristic level of anxiety and then distributed randomly to the different groups in the experiment. In the memory study, "anxiety level" would be called a **matching variable**. Individuals in the memory experiment would be given a valid and reliable measure of anxiety, those with similar scores would be paired, and one person in each pair would be randomly assigned to the group getting the 2-second rate and the other the group with the 4-second rate. As an illustration of exactly how to accomplish matching in our hypothetical two-group experiment on memory, with anxiety as a matching variable, you should work through the example in Table 6.1

Matching sometimes is used when the number of subjects is small and random assignment alone is therefore risky and might not yield equivalent groups. Fung and Leung (2014), for instance, attempted to increase the social responsiveness of children with autism by exposing them to therapy dogs (i.e., golden retrievers in this case, dogs known be calm and friendly to children). They only had 10 children with autism in their study, however, so they used matching to create a therapy group and a non-therapy control group, with 5 children in each group. Their matching variables were intellectual ability and verbal fluency. Even with matching, however, their small sample size prevented them from finding much evidence for the effectiveness of the therapy.

In order to undertake matching, regardless of whether sample size is an issue, two important conditions must be met. First, you must have good reason to believe the matching variable will have a predictable effect on the outcome of the study—that is, you must be confident that the matching variable would be correlated with the dependent variable. Because you haven't run your study yet to know if your matching variable correlates with your dependent variable, you would usually determine this by closely reading and evaluating previous, related research. When the correlation between the matching variable and the dependent variable is high, the statistical techniques for evaluating matched-groups designs are sensitive to differences between the groups. On the other hand, if matching is done when the correlation between the matching variable and the dependent variable is low, the chance of finding a true difference between the groups declines. So it is important to be careful when selecting matching variables.

A second important condition for matching is that there must be a reasonable way of measuring or identifying participants on the matching variable. In some studies, participants must be tested on the matching variable first, then assigned to groups, and then put through the experimental procedure. Depending on the circumstances, this might require bringing participants into the lab on two separate occasions, which can create logistical problems. Also, the initial testing on the matching variable might give participants an indication of the study's purpose, thereby introducing bias into the study. The simplest matching situations occur when the matching variables are constructs that can be determined without directly testing the

Table 6.1 How to Use a Matching Procedure

In our hypothetical study on the effect of presentation rate on memory, suppose that the researcher believes that matching is needed. That is, the researcher thinks that anxiety might correlate with memory performance. While screening subjects for the experiment, then, the researcher gives potential subjects a reliable and valid test designed to measure someone's characteristic levels of anxiety. For the sake of illustration, assume that scores on this test range from 10 to 50, with higher scores indicating greater levels of typical anxiety felt by people. Thus, a matching procedure is chosen, in order to ensure that the two groups of subjects in the memory experiment are equivalent to each other in terms of typical anxiety levels.

Step 1. Get a score for each person on the matching variable. This will be their score on the anxiety test (ranging from 10 to 50). Suppose there will be 10 subjects ("Ss") in the study, 5 per group. Here are their anxiety scores:

S1: 32	S6: 45
S2: 18	S7: 26
S3: 43	S8: 29
S4: 39	S9: 31
S5: 19	S10: 41

Step 2. Arrange the anxiety scores in ascending order.

S2: 18	S1: 32
S5: 19	S4: 39
S7: 26	S10: 41
S8: 29	S3: 43
S9: 31	S6: 45

Step 3. Create five pairs of scores, with each pair consisting of quantitatively adjacent anxiety scores:

Pair 1:	18 and 19
Pair 2:	26 and 29
Pair 3:	31 and 32
Pair 4:	39 and 41
Pair 5:	43 and 45

Step 4. For each pair, randomly assign one subject to group 1 (2-sec/item) and one to group 2 (4-sec/item). Here's one possible outcome:

	2-Sec/Item Group	**4-Sec/Item Group**
	18	19
	29	26
	31	32
	39	41
	45	43
Mean anxiety	**32.4**	**32.2**

Now the study can proceed with some assurance that the two groups are equivalent (32.4 is virtually the same as 32.2) in terms of anxiety.

Note: If more than two groups are being tested in the experiment, the matching procedure is the same up to and including step 2. In step 3, instead of creating pairs of scores, the researcher creates clusters equal to the number of groups needed. Then in step 4, the subjects in each cluster are randomly assigned to the multiple groups.

participants (e.g., intellectual ability and verbal fluency for the children with autism, based on already-available information), or by matching on the dependent variable itself. For instance, in a memory study, participants could be given an initial memory test, then matched on their performance, and then randomly assigned to groups. Their preexisting memory ability would thereby be under control, and the differences in performance could be attributed to the independent variable.

A nice example of a study using a matching procedure examined the "testing effect" that you learned about in Research Example 2 in the Chapter 3 discussion of "what's next" thinking. Goosens, Camp, Verkoeijen, Tabbers, and Zwann (2014) completed a conceptual replication of the testing effect to see if it would apply to 9-year-old school children (as it had with college-aged students). Children learned vocabulary words either by a simple study procedure (similar to a typical laboratory procedure), an elaborative study procedure (similar to the kinds of classroom exercises used with school children), or a "retrieval practice" procedure that involved repeated testing. Sample size was not a concern, but the researchers believed that it was important to assign children to the three groups after matching them on their pre-existing vocabulary capability. That is, in terms of the two critical conditions we just described that make matching a good idea, (a) there was good reason to believe that children with advanced vocabulary skills would perform better in the study than those with weaker skills, and (b) it was not difficult to determine the vocabulary competence of the children before the study began. With the matching procedure insuring equivalent groups, the research continued, and the general finding was that retrieval practice was superior to the other two forms of studying, thus providing more evidence for the validity of the testing effect. You might also note that because the study was completed in a school environment instead of a laboratory, it enhanced the *ecological validity* of the testing effect.

One final point about matching is the practical difficulty of deciding how many matching variables to use and which are the best to use. In a testing effect study, should the children have also been matched on general intelligence? What about anxiety level? You can see that some judgment is required here, for matching is difficult to accomplish with more than one matching variable and sometimes results in eliminating participants because close matches cannot be made. The problem of deciding on and measuring matching variables is one reason research psychologists generally prefer to make the effort to recruit enough volunteers to use random assignment even when they might suspect that some extraneous variable correlates with the dependent variable. In memory research, for instance, researchers are seldom concerned about such extraneous factors as anxiety level, intelligence, or education level. They simply make the groups large enough and assume that random assignment will distribute these extraneous (and potentially confounding) factors evenly throughout the conditions of the study.

Self Test 6.1

1. What is the defining feature of a between-subjects design? What is the main control problem that must be solved with this type of design?
2. It is sometimes the case that a study using deception requires a between-subjects design. Why?
3. Sal wishes to see if the font used when printing a document will influence comprehension of the material in the document. He thinks about matching on verbal fluency. What two conditions must be in effect before this matching can occur?

Within-Subjects Designs

As mentioned at the start of this chapter, each participant is exposed to each level of the independent variable in a within-subjects design. Because everyone in this type of study is measured several times, this procedure is sometimes described as a *repeated-measures* design. One practical advantage of this design should be obvious: Fewer people need to be recruited. If you have a study comparing two experimental conditions and you want to test 20 people in Condition 1, you'll need to recruit 40 people for a between-subjects study, but only 20 for a within-subjects study.

Within-subjects designs are sometimes the only reasonable choice. In experiments in areas such as physiological psychology and sensation and perception, comparisons often are made between conditions that require just a brief time to test but might demand extensive preparation. For example, a perceptual study using the Müller-Lyer illusion might vary the orientations of the lines to see if the illusion is especially strong when presented vertically (see Figure 6.1). The task might involve showing the illusion on a computer screen and asking the participant to tap a key that gradually changes the length of one of the lines. Participants are told to adjust the line until both lines are perceived to be the same length. Any one trial might take no more than 5 seconds, so it would be absurd to make the illusion orientation variable a between-subjects factor and use one person for a fraction of a minute. Instead, it makes more sense to make the orientation variable a within-subjects factor and give each participant a sequence of trials to cover all levels of the variable and probably duplicate each level several times (to get consistent measurements). And unlike the attractive/unattractive Barbara Helm study, serving in one condition in this perception study would not make it impossible to serve in another.

A within-subjects design might also be necessary when volunteers are scarce because the entire population of interest is small. Studying astronauts or people with special expertise (e.g., world-class chess players, to use an example you will see shortly) are just two examples. Of course, there are times when, even with a limited population, the design may require a between-subjects manipulation. Fung and Leung (2014), mentioned earlier, evaluated the effects of a pet-assisted therapy for children with autism in an experiment that required comparing those in therapy with others in a control group not being exposed to the therapy.

Besides convenience, another advantage of within-subjects designs is that they eliminate problems associated with creating equivalent groups associated with between-subjects designs. Having different individuals in each condition in a between-subjects design introduces more *variability* in each condition of your study. Even with random assignment, a portion of the *variance* in a between-subjects design can result from individual differences between subjects in the different groups. But in a within-subjects design, any between-condition individual difference variance disappears. Let's look at a simple example.

Suppose you are comparing two golf balls for distance. You recruit 10 professional golfers and randomly assign them to two groups of 5. After loosening up, each golfer hits one ball or the other. Here are the results (each number refers to the number of yards a ball has been hit).

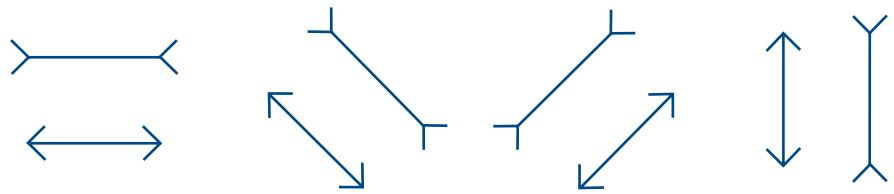

FIGURE 6.1
Set of four Müller-Lyer illusions: horizontal, 45° left, 45° right, vertical.

Pros in the First Group	*Golf ball 1*	*Pros in the Second Group*	*Golf ball 2*
Pro 1	255	Pro 6	269
Pro 2	261	Pro 7	266
Pro 3	248	Pro 8	260
Pro 4	256	Pro 9	273
Pro 5	245	Pro 10	257
M	**253.00**	***M***	**265.00**
SD	**6.44**	***SD***	**6.52**

Note several points here. First, there is variability within each group, as reflected in the standard deviation for each (6.44 yards and 6.52 yards). Second, there is apparently an overall difference between the groups (253 yards and 265 yards). The pros in the second group hit their ball farther than the pros in the first group. Why? Three possibilities:

a. *The golf ball:* Perhaps the brand of golf ball hit by the second group simply goes farther (this, of course, is the research hypothesis).

b. *Individual differences:* Maybe the golfers in the second group are stronger or more skilled than those in the first group.

c. *Chance:* Perhaps this is not a statistically significant difference, and even if it is, there's a 5% chance it is a Type I error if the null hypothesis (i.e., no real difference) is actually true.

The chances that the second possibility is a major problem are reduced by the procedures for creating equivalent groups described earlier. Using random assignment or matching allows you to be reasonably sure the second group of golfers is approximately equal to the first group in ability, strength, and so on. Despite that, however, it is still possible that *some* of the difference between these groups can be traced to the individual differences between the groups. This problem simply does not occur in a within-subjects design, however. Suppose you repeated the study but used just the first five golfers, and each pro hit ball 1, and then ball 2. Now the table looks like this.

Pros in the First Group	*Golf ball 1*	*Golf ball 2*
Pro 1	255	269
Pro 2	261	266
Pro 3	248	260
Pro 4	256	273
Pro 5	245	257
M	**253.00**	**265.00**
SD	**6.44**	**6.52**

Of the three possible explanations for the differences in the first set of data, when there were two groups, explanation b. can be eliminated for the second set. In the first set, the difference in the first row between the 255 yards and the 269 yards could be due to the difference between the balls, or individual differences between pros 1 and 6, 2 and 7 or chance. In the second set of data, there is no second group of golfers, so the second possibility is gone. Thus, in a within-subjects design, individual differences are eliminated from the estimate of the variability between conditions. Statistically, this means that, in a within-subjects design, an *inferential analysis* will

be more sensitive to small differences between means than it would in a between-subjects design. We will describe in more depth inferential analysis in Chapter 7 when we describe our first inferential statistic, the *t*-test.

But wait. Are you completely satisfied that in the second case, the differences between the first set of scores and the second set could be due *only* to chance factors and/or the superiority of the second ball? Are you thinking that perhaps pro 1 actually changed in some way between hitting ball 1 and hitting ball 2? Although it's unlikely that the golfer will add 10 pounds of muscle between swings, what if some kind of practice or warm-up effect was operating? Or perhaps the pro detected a slight malfunction in his swing at ball 1 and corrected it for ball 2. Or perhaps the wind changed. In short, with a within-subjects design, a major problem is that once a participant has completed the first part of a study, the experience or altered circumstances could influence performance in later parts of the study. The problem is referred to as a sequence or **order effect**, and it can operate in several ways.

First, Trial 1 might affect the participant so performance on Trial 2 is steadily improved, as in the example of a practice effect. On the other hand, sometimes repeated trials produce gradual fatigue or boredom, and performance steadily declines from trial to trial. These two effects can both be referred to as **progressive effects** because it is assumed that performance changes steadily (progressively) from trial to trial. Second, some sequences might produce effects different from those of other sequences, what could be called a **carry-over effect**. Thus, in a study with two basic conditions, experiencing the first condition before the second might affect the person much differently than experiencing the second before the first. For example, suppose you are studying the effects of noise on a problem-solving task using a within-subjects design. Let's say participants will be trying to solve anagram problems (rearrange letters to form words). In condition UPN (UnPredictable Noise), they have to solve the anagrams while distracting noises come from the next room, and these noises are presented randomly and therefore are unpredictable. In condition PN (Predictable Noise), the same total amount of noise occurs; however, it is not randomly presented but instead occurs in predictable patterns (e.g., every 30 seconds). If you put the people in condition UPN first, and then in PN, they will probably do poorly in UPN (most people do). This poor performance might discourage them and *carry over* to condition PN. They should do better in PN, but as soon as the noise begins, they might say to themselves, "Here we go again," and perhaps not try as hard. On the other hand, if you run condition PN first, the predictable noise, your subjects might do reasonably well (most people do), and some of the confidence might carry over to the second part of the study—UPN. ("Here comes the noise again, but I handled it before, I can handle it again.") When they encounter condition UPN in the second part of the study, they might do better than you would ordinarily expect. Thus, performance in condition UPN might be much worse in the sequence UPN–PN than in the sequence PN–UPN. Furthermore, a similar problem would occur for condition PN. In short, the order in which the conditions are presented, independently of practice or fatigue effects, might influence the study's outcome. In studies where carryover effects might be suspected, researchers usually decide to use a between-subjects rather than a within-subjects design. Indeed, studies comparing predictable and unpredictable noise typically put people in two different groups precisely because of this carryover issue.

Controlling Order Effects

The typical way to control order effects in a within-subjects design is to use more than one sequence, a strategy known as **counterbalancing**. As is clear from the predictable versus unpredictable noise example, the procedure works better for progressive effects than for carryover effects. There are two general categories of counterbalancing, depending on whether participants are tested in each experimental condition just one time or are tested more than once per condition.

Testing Once per Condition

In some experiments, participants are tested in each of the conditions but only once per condition. Consider, for example, an interesting study by Reynolds (1992) on the ability of chess players to recognize the level of expertise in other chess players. He recruited 15 chess players with different degrees of expertise from various clubs in New York City and asked them to look at 6 chess games that were said to be in progress (i.e., about 20 moves into the game). On each trial, the players examined the board of an in-progress game (they were told to assume the two players in each game were of equal ability) and estimated the skill level of the players according to a standard rating system. The games were deliberately set up to reflect different levels of player expertise. Reynolds found the more highly skilled of the 15 chess players made more accurate estimates of the ability reflected in the board setups they examined than did the less skilled players.

Can you see that the Reynolds study used within-subjects design? Each of the 15 participants examined all six games. Also, you can see it made sense for each game to be evaluated just one time by each player. Hence, Reynolds was faced with the question of how to control for order effects that might be present. He certainly didn't want all 15 participants to see the 6 games in exactly the same order. How might he have proceeded?

Complete Counterbalancing

Whenever participants are tested once per condition in a within-subjects design, one solution to the order problem is to use **complete counterbalancing**. This means every possible sequence will be used at least once. The total number of sequences needed can be determined by calculating $X!$, where X is the number of conditions, and ! stands for the mathematical calculation of a factorial. For example, if a study has three conditions, there are six possible orders that can be used:

$$3! = 3 \times 2 \times 1 = 6$$

The six sequences in a study with conditions A, B, and C would be:

A B C	B A C
A C B	C A B
B C A	C B A

Then, subjects in this study would be randomly assigned to one of the six orders.

The problem with complete counterbalancing is that as the number of levels of the independent variable increases, and the possible orders needed increase dramatically. Six orders are needed for three conditions, but look what happens if you add a fourth condition:

$$4! = 4 \times 3 \times 2 \times 1 = 24$$

By adding just one more condition, the number of possible orders increases from 6 to 24. As you can guess, complete counterbalancing was not possible in the Reynolds study unless he recruited many more than 15 chess players. In fact, with 6 different games (i.e., conditions), he would need to find 6! or 720 players to cover all of the possible orders. Clearly, Reynolds used a different strategy.

Partial Counterbalancing

Whenever a subset of the total number of orders is used, the result is called **partial counterbalancing** or, sometimes, *incomplete counterbalancing* This can be accomplished by taking a random sample of orders from the complete set of all possible orders or, more simply, by randomizing

the order of conditions for each subject.[1] The latter was Reynolds's solution—"the order of presentation [was] randomized for each subject" (Reynolds, 1992, p. 411). Sampling from the population of orders is a common strategy whenever there are fewer participants available than possible orders or when there is a fairly large number of conditions.

Reynolds (1992) sampled from the total number of orders, but he could have chosen another approach that is used sometimes: the balanced **Latin square**. This device gets its name from an ancient Roman puzzle about arranging Latin letters in a matrix so each letter appears only once in each row and once in each column (Kirk, 1968). The Latin square strategy is more sophisticated than choosing a random subset of the whole. With a perfectly balanced Latin square, you are assured that (a) every condition of the study occurs equally often in every sequential position, and (b) every condition precedes and follows every other condition exactly once. Also, the number of rows in a Latin square is exactly equal to the number of elements in the study that are in need of counterbalancing. Here is an example of a 6x6 square.[2] Think of each letter as one of the six games inspected by Reynolds's chess players.

A	B	F	C	E	D
B	C	**A**	D	F	E
C	D	B	E	**A**	F
D	E	C	F	B	**A**
E	F	D	**A**	C	B
F	**A**	E	B	D	C

We've boldfaced condition **A** (representing chess game setup #1) to show you how the square meets the two requirements listed in the preceding paragraph. First, condition A occurs in each of the six sequential positions (first in the first row, third in the second row, etc.). Second, A is followed by each of the other letters exactly once. From the top to the bottom rows, (1) A is followed by B, D, F, nothing, C, and E, and (2) A is preceded by nothing, C, E, B, D, and F. The same is true for each of the other letters. Once you have generated a balanced Latin square, you can then randomly assign participants to one of the orders (rows). In our example, participants would be randomly assigned to one of the six orders, represented as rows in the Latin square.

When using Latin squares, it is important for the number of subjects in the study to be equal to or a multiple of the number of rows in the square. That Reynolds had 15 subjects in his study tells you he didn't use a Latin square. If he had added three more chess players, giving him an *N* of 18, he could have randomly assigned three players to each of the 6 rows of the square (3 x 6 = 18).

Testing More than Once per Condition

In the Reynolds (1992) study, it made no sense to ask the chess players to look at any of the six games more than once. Similarly, if participants in a memory experiment are asked to study and recall four lists of words, with the order of the lists determined by a 4 x 4 Latin square, they are seldom asked to study and recall any particular list a second time unless the researcher is specifically interested in the effects of repeated trials on memory. However, in many studies, it is reasonable, even necessary, for participants to experience each condition more than once. This often happens in research in perception and attention, for instance. A look back at the Müller-Lyer illusions in Figure 6.1 provides an example.

[1] Strictly speaking, these two procedures are not the same. Sampling from all possible orders guarantees that no one order will ever be repeated; randomizing the order for each subject does not carry that guarantee.

[2] There are easy-to-use formulas available for building Latin squares; instructions on how to build them and examples of Latin squares can be found quickly with a Google search.

Suppose you were conducting a study in which you wanted to see if participants would be more affected by the Müller-Lyer illusion when it was presented vertically than when shown horizontally or at a 45° angle. Four conditions of the study shown in Figure 6.1, when assigned to the letters A, B, C, and D, are as follows:

A = horizontal
B = 45° to the left
C = 45° to the right
D = vertical

Participants in the study are shown the illusion on a computer screen and make adjustments to the lengths of the parallel lines until they perceive the lines to be equal. A researcher would probably want to test participants in each of the four conditions (A, B, C, and D) more than once to get a more reliable measure of perception and to reduce any chances of measurement error. The four conditions could be presented multiple times to people according to one of two basic procedures.

Reverse Counterbalancing

When using **reverse counterbalancing**, the experimenter simply presents the conditions in one order and then presents them again in the reverse order. In the illusion case, the order would be A–B–C–D, then D–C–B–A. If the researcher wants the participant to perform the task more than twice per condition, and this is common in perception research, this sequence of orders could be repeated as many times as necessary. Hence, if you wanted each participant to adjust each of the four illusions of Figure 6.1 six separate times, and you decided to use reverse counterbalancing, half the participants would be randomly assigned to see the illusions in this order:

A-B-C-D – D-C-B-A – A-B-C-D – D-C-B-A – A-B-C-D – D-C-B-A

while the remaining would see this order:

D-C-B-A – A-B-C-D – D-C-B-A – A-B-C-D – D-C-B-A – A-B-C-D

Reverse counterbalancing was used in one of psychology's most famous studies, completed in the 1930s by J. Ridley Stroop. You've probably tried the Stroop task yourself—when shown color names printed in the wrong colors, you were asked to name the color rather than read the word. That is, when shown the word *red* printed in blue ink, the correct response is "blue," not "red." Stroop's study is a classic example of a particular type of design described in the next chapter, so you will be learning more about his work when you encounter Box 7.1 in Chapter 7.[3]

Block Randomization

A second way to present a sequence of conditions when each condition is presented more than once is to use **block randomization**, where the basic rule is that every condition must occur once before any condition can be repeated. Within each block, the order of conditions is randomized. This strategy eliminates the possibility that participants can predict what is coming next, a problem that can occur with reverse counterbalancing, especially if the reversals occur

[3] Although reverse counterbalancing normally occurs when participants are tested more than once per condition, the principle can also be applied in a within-subjects design in which participants see each condition only once. Thus, if a within-subjects study has six conditions, each tested only once per person, half of the participants could get the sequence A-B-C-D-E-F, while the remaining participants experience the reverse order (F-E-D-C-B-A).

several times. In principle, this is the same procedure outlined earlier in the context of how to assign subjects randomly to groups in a between-subjects experiment. Given the ease of obtaining computer-generated randomized orders, this procedure is used more frequently than reverse counterbalancing.

Using the illusions example again (Figure 6.1), participants would encounter all four conditions in a randomized order, then all four again but in a block with a new randomized order, and so on, for as many blocks of four as needed. A reverse counterbalancing would look like this:

A-B-C-D – D-C-B-A

A block randomization procedure might produce either of these two sequences (among others):

B-C-D-A – C-A-D-B or C-A-B-D – A-B-D-C

To give you a sense of how block randomization works in an actual within-subjects experiment employing many trials, consider the following study, which suggests that wearing red in an aggressive sport might help you win.

Research Example 8—Counterbalancing with Block Randomization

An interesting study by Hagemann, Strauss, and Leißing (2008) suggests that those who referee combative forms of athletics (e.g., wrestling, judo) can be influenced in their scoring by the colors worn by combatants. They asked 42 experienced tae kwon do referees to examine video segments of matches between two males of equal ability and to assign scoring points that followed their standard set of rules. Each of the combatants wore a white tae kwon do uniform, but each also wore "trunk and head protectors" (p. 769) that were either red or blue.

Each video segment lasted 4.4 seconds and 11 different videos were made. Each referee saw all of the videos, making this a within-subjects design requiring counterbalancing. Block randomization was set up this way: Each referee saw a block of the 11 videos in a random order, and then saw a second block of the same 11 videos, also in a random order. Why two blocks? In one of the blocks of 11 videos, one of the fighters wore red protective gear while the other wore blue. In the second block, the same 11 clips were seen, but the researchers digitally switched the colors of the protective gear. The two blocks were also counterbalanced, so half the refs saw each block first. Thus, referees saw the same tae kwon do event twice, with players dressed in different colors. By creating this design, Hagemann et al. (2008) could determine it was more likely their independent variable (color of the protectors) that would cause differences in referees' judgments rather than the fighters themselves or the fight events themselves.

For experienced and highly competent referees, the color of protective gear should make no difference in judging the quality of the fighting performance, but it did. Figure 6.2 shows the results. The two bars on the left show that in the first block of 11 videos, those wearing red protective gear scored higher than those wearing blue. When the colors were reversed (two bars on the right), even though the *same events* were seen, red protective gear still gave a perceived advantage.

This outcome, of course, makes little sense if you believe referees are objective, and Hagemann et al. (2008) did not provide much of an explanation for their results. Other studies have proposed that red suggests aggressiveness and blue projects calmness, though, so it is conceivable those associations played a role in the biased judgments.

One final point: Notice that Figure 6.2 includes vertical lines at the top of each bar (they also appear in Figure 5.1 in Chapter 5). These are called **error bars**, and they indicate the amount of variability that occurred within each condition. In this case, the distance between the top of a bar and the end of the vertical line represents standard errors, a concept related to the standard

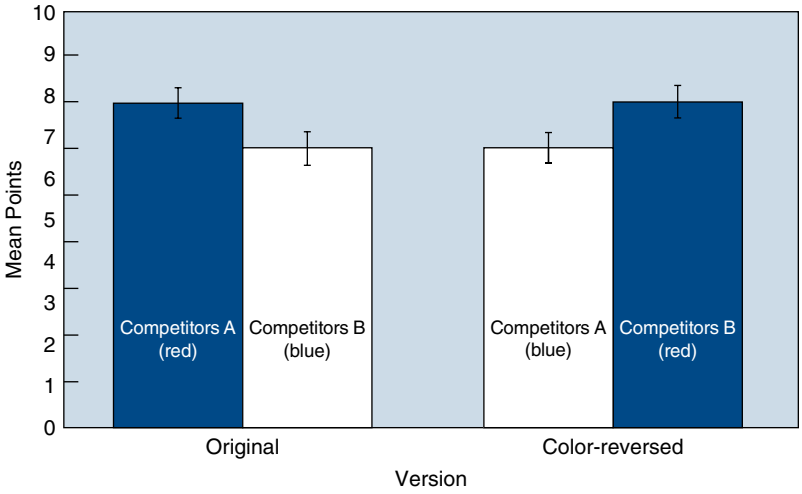

FIGURE 6.2
The effects of color bias on referee judgment of tae kwon do matches (from Hagemann, Strauss, & Leißing (2008).

deviation (error bars can sometimes mean other things, such as confidence intervals). Including error bars in graphs enables the researcher to show the reader both central tendency *and* variability; these bars are now standard features of graphical displays of data.

SELF TEST

6.2

1. What is the defining feature of a within-subjects design? What is the main control problem that must be solved with this type of design?
2. If your IV has six levels, each tested just once per subject, why are you more likely to use partial counterbalancing than complete counterbalancing?
3. If participants are going to be tested more than one time for each level of the IV, what two forms of counterbalancing may be used?

Methodological Control in Developmental Research

As you have learned, the researcher must weigh several factors when deciding whether to use a between-subjects design or a within-subjects design. Additional considerations affect researchers in developmental psychology, where two specific versions of these designs occur. These methods are known as *cross-sectional* and *longitudinal designs*.

You've seen these terms before if you have taken a course in developmental or child psychology. Research in these areas includes age as the prime independent variable; after all, the name of the game in developmental psychology is to discover how we change as we grow older. A **cross-sectional study** takes a between-subjects approach. A cross-sectional study comparing the language performance of 3-, 4-, and 5-year-old children would use three groups of children. A **longitudinal study**, on the other hand, takes a within-subjects or repeated measures approach in which a single group of subjects is studied over time. The same language

study would measure language behavior in a group of 3-year-olds and then study these same children when they turned 4, and again at age 5.

The obvious advantage of the cross-sectional approach to the experiment on language is time; data collection for a study comparing 3-, 4-, and 5-year-olds might take a month. If done as a longitudinal study, data collection would take at least 2 years. However, a potentially serious difficulty with some cross-sectional studies is a special form of the problem of nonequivalent groups and involves **cohort effects**. A cohort is a group of people born at about the same time. If you are studying three age groups, they differ not only in chronological age but also in terms of the environments in which they were raised. The problem is not especially noticeable when comparing 3-, 4-, and 5-year-olds, but what if you're interested in whether intelligence declines with age and decide to compare groups aged 45, 65, and 85? You might indeed find a decline with age but does it mean that intelligence decreases with age, or might the differences relate to the very different life histories of the three groups? For example, the 85-year-olds went to school during the Great Depression, the 65-year-olds were educated during the post–World War II boom, and the 45-year-olds were raised on TV. These factors could bias the results. Indeed, this outcome has occurred. Early research on the effects of age on IQ suggested that significant declines occurred, but these studies were cross sectional (e.g., Miles, 1933). Subsequent longitudinal studies revealed a different pattern, however (Schaie, 1988). For example, verbal abilities show minimal decline, especially if the person remains verbally active (moral: for some skills at least, use it or lose it).

While cohort effects can plague cross-sectional studies, longitudinal studies also have problems, most notably with *attrition* (Chapter 5). If a large number of participants drop out of the study, the group completing it may be different from the group starting it. Referring to the age and IQ example, if people stay healthy, they may remain more active intellectually than if they are sick all of the time. If they are chronically ill, they may die before a study is completed, leaving a group that may be generally more intelligent than the group starting the study. Longitudinal studies also pose potential ethical problems. As people develop and mature, they might change their attitudes about their willingness to participate. Most researchers doing longitudinal research recognize that informed consent is an ongoing process, not a one-time event. Ethically sensitive researchers will periodically renew the consent process in long-term studies, perhaps every few years (Fischman, 2000).

In trying to balance cohort and attrition problems, many developmental researchers use a strategy that combines cross sectional with longitudinal studies; one such design is called a **cohort sequential design**. In such a study, a group of subjects is selected and retested every few years, and additional cohorts are selected every few years and also retested over time. So different cohorts are continually being retested. To take a simple example, suppose you wished to examine the effects of aging on memory, comparing ages 55, 60, and 65. In the study's first year, you would recruit a group of 55 year olds. Then, every 5 years after that, you would recruit new groups of 55-year-olds *and* retest those who had been recruited earlier. Schematically, the design for a study that began in the year 2010 and lasted for 30 years would look like this (the numbers in the matrix refer to the age of the subjects at any given testing point):

	Year of the Study						
Cohort #	*2010*	*2015*	*2020*	*2025*	*2030*	*2035*	*2040*
1	55	60	65				
2		55	60	65			
3			55	60	65		
4				55	60	65	
5					55	60	65

So in 2010, you test a group of 55 year olds. In 2015, you retest these same people (now 60 years old), along with a new group of 55-year-olds. By year three (2020), you have cohorts for all three age groups. By 2040, combining the data in each of the diagonals will give you an overall comparison of those aged 55, 60, and 65. Comparing the data in the rows gives you longitudinal designs, while comparing data in columns (especially 2020, 2025, and 2030) gives you cross-sectional comparisons. Comparing the rows enables a comparison of overall differences among cohorts. In actual practice, these designs are more complicated because researchers will typically start the first year of the study with a fuller range of ages. But the diagram gives you the basic idea.

Perhaps the best-known example of this type of sequential design is a long series of studies by K. Warner Schaie (2005), known as the Seattle Longitudinal Study. Begun in 1956, it was designed to examine age-related changes in various mental abilities. The initial cohort had 500 people, ranging in age from their early 20's to their late 60's (as of 2005, 38 of these subjects were still in the study, 49 years later!). The study has added a new cohort at 7-year intervals ever since 1956 and has recently reached the 50-year mark. About 6,000 in all people have participated. In general, Schaie and his team have found that performance on mental ability tasks declines slightly with age, but with no serious losses before age 60, and the losses can be reduced by good physical health and lots of crossword puzzles. Concerning cohort effects, they have found that overall performance has been progressively better for those born more recently. Presumably, those born later in the 20th century have had the advantages of better education, better nutrition, and so on.

The length of Schaie's (2005) Seattle project is impressive, but the world record for perseverance in a repeated-measures study occurred in what is arguably the most famous longitudinal study of all time. Before continuing, read Box 6.1, which chronicles the epic tale of Lewis Terman's study of gifted children.

BOX 6.1 CLASSIC STUDIES—The Record for Repeated Measures

In 1921, the psychologist Lewis Terman (1877–1956) began what became the longest-running repeated-measures design in the history of psychology. A precocious child himself, Terman developed an interest in studying gifted children as a graduate student. His doctoral dissertation, supervised by Edmund Sanford at Clark University in 1905, was his first serious investigation of giftedness; in it, he compared what he labeled "bright" and "dull" local school children to see which tests might best distinguish between them (Minton, 1987). This early interest in giftedness and mental testing foreshadowed Terman's two main contributions to psychology. First, he transformed the intelligence test created by Alfred Binet of France into the popular Stanford-Binet IQ test. Second, he began a longitudinal study of gifted children that continued long after he died.

Terman was motivated by the belief, shared by most mental testers of his day, that the United States should become a meritocracy—that is, he believed that positions of leadership should be held by those most *able* to lead. You can see how this belief led to his interests in IQ and giftedness. To bring about a meritocracy, there must be ways to recognize (i.e., measure) and nurture talent.

Unlike his dissertation, which studied just 14 children, Terman's longitudinal study of gifted children was a mammoth undertaking. Through a variety of screening procedures, he recruited 1,470 children (824 boys and 646 girls). Most were in elementary school, but a group of 444 were in junior or senior high school (sample numbers from Minton, 1988). Their average IQ score was 150, which put the group roughly in the top 1 percent of the population. (Despite the large sample size it is worth noting that the sample was severely biased—heavily weighted with White, middle- and upper-class children.) Each child selected was given an extensive battery of tests and questionnaires by the team of graduate students assembled by Terman. By the time the initial testing was complete, each child had a file of about 100 pages long (Minton, 1988)! The results of the first analysis of the group were published in more than 600 pages as the *Mental and Physical Traits of a Thousand Gifted Children* (Terman, 1925).

Terman intended to do just a single brief follow-up study, but the project took on a life of its own. The sample was retested in the late 1920s (Burks, Jensen, & Terman, 1930), and additional follow-up studies during Terman's lifetime were published 25 (Terman & Oden, 1947) and 35 (Terman & Oden, 1959) years after the initial testing. Following Terman's death, the project was taken over by Robert Sears, a member of the gifted group and a well-known psychologist in his own right. In the foreword to the 35-year follow-up, Sears wrote: "On actuarial grounds, there is considerable likelihood that the last of Terman's Gifted Children will not have yielded his last report to the files before the year 2010!" (Terman & Oden, 1959, p. 9). Between 1960 and 1986, Sears produced five additional follow-up studies of the group, and he was working on a book-length study of the group as they aged when he died in 1989 (Cronbach, Hastorf, Hilgard, & Maccoby, 1990). The book was eventually published as *The Gifted Group in Later Maturity* (Holahan, Sears, & Cronbach, 1995).

Three points are worth making about this mega-longitudinal study. First, Terman's work questioned the stereotype of the gifted child as someone who was brilliant but socially inept and prone to burnout early in life. Rather, the members of his group as a whole were both brilliant and well-adjusted, and they became successful as they matured. By the time they reached maturity, "the group had produced thousands of scientific papers, 60 nonfiction books, 33 novels, 375 short stories, 230 patents, and numerous radio and television shows, works of art, and musical compositions" (Hothersall, 1990, p. 353). Second, the data collected by Terman's team continues to be a source of rich archival information for modern researchers (look ahead to Chapter 10 for more on archival research methodology). For instance, studies have been published on the careers of the gifted females in Terman's group (Tomlinson-Keasy, 1990) and on the predictors of longevity in the group (Friedman et al., 1995). Third, Terman's follow-up studies are incredible from the methodological standpoint of a longitudinal study's typical nemesis: *attrition*. The following figures (taken from Minton, 1988) are the percentage of living participants who participated in the first three follow-ups:

After 10 years: 92%

After 25 years: 98%

After 35 years: 93%

These are remarkably high numbers and reflect the intense loyalty Terman and his group had for each other. Members of the group referred to themselves as "Termites," and some even wore termite jewelry (Hothersall, 1990). Terman corresponded with hundreds of his participants and genuinely cared for them. After all, the group represented the type of person Terman believed held the key to America's future.

Controlling for the Effects of Bias

Because humans are always the experimenters and usually the participants in psychology research, there is the chance the results of a study could be influenced by some human bias, a preconceived expectation about what is to happen in an experiment. These biases take several forms but fall into two broad categories: those affecting experimenters and those affecting research participants. These two forms of bias often interact.

Experimenter Bias

As well as illustrating falsification and parsimony, the Clever Hans case (Box 3.2 in Chapter 3) is often used to show the effects of **experimenter bias** on the outcome of some study. Hans's questioner, knowing the outcome to the question "What is 3 times 3?" sent subtle head movement cues that were read by the apparently intelligent horse. Similarly, experimenters testing hypotheses sometimes may inadvertently do something that leads participants to behave in ways that confirm the hypothesis. Although the stereotype of the scientist is that of an objective, dispassionate, even mechanical person, the truth is that researchers can become emotionally involved in their research. It's not difficult to see how a desire to confirm a strongly held hypothesis might

lead an unwary but emotionally involved experimenter to behave (without awareness) in such a way as to influence the outcome of the study.

For one thing, biased experimenters might treat the research participants in the various conditions differently. Rosenthal developed one procedure demonstrating this. Participants in one of his studies (e.g., Rosenthal & Fode, 1963a) were shown a set of photographs of faces and asked to make judgments about the people pictured in them. For example, they might be asked to rate each photo on how successful the person seemed to be, with the interval scale ranging from −10 (*total failure*) to +10 (*total success*). All participants saw the same photos and made the same judgments. The independent variable was experimenter expectancy. Some experimenters were led to believe most subjects would give people the benefit of the doubt and rate the pictures positively; other experimenters were told to expect negative ratings. Interestingly, the experimenter's expectancies typically produced effects on the subjects' rating behavior, even though the pictures were identical for both groups. How can this be?

According to Rosenthal (1966), experimenters can innocently communicate their expectancies in a number of subtle ways. For instance, on the person perception task, the experimenter holds up a picture while the participant rates it. If the experimenter is expecting a +8 and the person says –3, how might the experimenter act—with a slight frown perhaps? How might the participant read the frown? Might he or she try a +7 on the next trial to see if this could elicit a smile or a nod from the experimenter? In general, could it be that experimenters in this situation, without being aware of it, subtly shape the response of their participants? Does this remind you of Clever Hans?

Rosenthal has even shown that experimenter expectancies can be communicated to subjects in animal research. For instance, he found that rats learned mazes faster for experimenters who *thought* their animals had been bred for maze-running ability than for those who expected their rats to be "maze-dull" (Rosenthal & Fode, 1963b). The rats, of course, were randomly assigned to the experimenters and equal in ability. The determining factor seemed to be that experimenters expecting their rats to be "maze-bright" treated them better; for example, they handled them more, a behavior known to affect learning.

Rosenthal's research has been criticized (e.g., Barber, 1976) on statistical (omitting some data, a *QRP* – see Chapter 3) and methodological grounds (real studies never have such large numbers of experimenters), but the experimenter expectancy effect cannot be ignored; it has been replicated in a variety of situations and by many researchers other than Rosenthal and his colleagues (e.g., Word, Zanna, & Cooper, 1974). Furthermore, experimenters can be shown to influence the outcomes of studies in ways other than through their expectations. The behavior of participants can be affected by the experimenter's race and gender, as well as by demeanor, friendliness, and overall attitude (Adair, 1973). An example of the latter is a study by Fraysse and Desprels-Fraysse (1990), who found that preschoolers' performance on a cognitive task could be influenced by experimenter attitude. The children performed significantly better with "caring" than with "indifferent" experimenters.

Controlling for Experimenter Bias

It is probably impossible to eliminate experimenter effects completely. Experimenters cannot be turned into machines. However, one strategy to reduce bias is to mechanize procedures as much as possible. For instance, it's not hard to remove a frowning or smiling experimenter from the person perception task. Instead, subjects can be shown photos on a screen and asked to make their responses with a key press while the experimenter is out of sight or in a different room entirely.

Similarly, procedures for testing animals automatically have been available since the 1920s, even to the extent of eliminating human handling completely. E. C. Tolman didn't wait for computers to come along before inventing "a self-recording maze with an automatic delivery table" (Tolman, Tryon, & Jeffries, 1929). The delivery table was so-called because it "automatically

delivers each rat into the entrance of the maze and 'collects' him at the end without the mediation of the experimenter. Objectivity of scoring is insured by the use of a device which automatically records his path through the maze" (Tryon, 1929, p. 73). Today such automation is routine. Furthermore, computers make it easy to present instructions and stimuli to participants while keeping track of data.

Experimenters can mechanize many procedures, to some degree at least, but the experimenter interacts with every participant nonetheless (e.g., during the consent process). Hence, it is important for experimenters to be trained in how to be experimenters and for the experiments to have highly detailed descriptions of the sequence of steps that experimenters should follow in every research session. These descriptions are called research **protocols**.

A common strategy for controlling for experimenter bias is to use what is called a **double blind** procedure. This means that experimenters are kept unaware, or "in the dark" (blind), about what to expect of participants in a particular testing session; in addition, subjects do not know what to expect. Specifically, neither the experimenters nor the participants know which condition is being tested on any particular trial—hence the designation *double*. (A **single blind** procedure is one in which subjects are unaware but the experimenters know the condition in which each subject is being tested.) A double blind procedure can be accomplished when the principal investigator sets up the experiment but a colleague (usually a graduate student) actually collects the data. Double blind studies are not always easy to create but can be more easily managed, as illustrated in a study by Williams and Bargh (2008). They wondered whether "experiences of physical warmth (or coldness) would increase feelings of interpersonal warmth (or coldness), without the person's being aware of this influence" (p. 606). In their procedure, experimenters handed subjects a cup of either hot or iced coffee. Those whose hands had been warmed subsequently judged a neutral third person as warmer in personality than those given cold hands. The researchers recognized the possibility for bias; the experimenters obviously knew who had received the warm coffee and who had received the iced coffee (i.e., a single blind procedure was in effect). To eliminate the chance that experimenters had "inadvertently treated participants in the two conditions differently" (p. 607), Williams and Bargh ran a second study using a double blind procedure; they created physical warmth and coldness in subjects in a way that kept experimenters uninformed about which condition was being tested. In essence, they replicated their results using single- and double-blind procedures—warm or cold hands influenced participants' behavior. This indicated that the effects of physical warmth on personality judgments was not due to experimenter bias but likely do to a real relationship between warmth and personality.

Research Example 9, which has a practical take-home message for senior citizens (have some coffee around 4:00 in the afternoon) illustrates the effective use of a double blind procedure.

Research Example 9—Using a Double Blind Procedure

There is considerable evidence that as we age, we become less efficient cognitively in the afternoon. Also, older adults are more likely to describe themselves as "morning persons." Ryan, Hatfield, and Hofstetter (2002) wondered if the cognitive decline, as the day wears on, could be neutralized by America's favorite drug—caffeine. They recruited 40 seniors, all 65 or older and self-described as (a) morning types and (b) moderate users of caffeine, and placed them in either a caffeine group or a decaf group (using Starbucks house blend). At each testing session, participants drank a 12-ounce cup of coffee, either caffeinated or not; 30 minutes later, they were given a standardized memory test. The second independent variable was time of testing—either 8:00 A.M. or 4:00 P.M. Subjects were tested twice, once at each time, with a 5- to 11-day interval between sessions. Thus, "time of testing" was a within-subjects manipulated independent variable; whether the seniors drank caffeine or decaf was a between-subjects manipulated independent variable.

The procedure was a double blind one because the experimenters administering the memory tests did not know which participants had ingested caffeine and the seniors did not know which

type of coffee they were drinking. And to test for the adequacy of the control procedures, the researchers completed a clever *manipulation check* (a concept you learned about in Chapter 3). At the end of the study, during debriefing, they asked the participants to guess whether they had been drinking the real stuff or the decaf. The accuracy of the seniors' responses was at chance level; they had no idea what they had been drinking. In fact, most guessed incorrectly that they had been given regular coffee during one testing session and decaf at the other. In fact, all subjects had been given either caffeinated coffee at both sessions or decaf at both sessions.

The researchers also did a nice job of incorporating some of the other control procedures you learned about in this chapter. For instance, the seniors were randomly assigned to the two groups, and this *random assignment* seemed to produce the desired equivalence; the groups were indistinguishable in terms of age, education level, and average daily intake of caffeine.[4] Also, *counterbalancing* was used to ensure half of the seniors were tested first in the morning, then the afternoon, while the other half were tested in the afternoon-then-morning sequence.

The results? Time of day did not seem to affect an immediate short-term memory task, but it had a significant effect on a more difficult longer-term memory task. For this second task, seniors studied some information, waited 20 minutes, tried to recall the information, and then completed a recognition test for that same information. Caffeine prevented the decline for this more demanding task. On both the delayed recall and the delayed recognition tasks, seniors scored equally well in the morning sessions. In the afternoon sessions, however, those ingesting caffeine still did well, but the performance of those taking decaf declined. On the delayed recall task, for instance, here are the means of the total number of words recalled (out of 16). Also, remember from Chapter 4 that, when reporting descriptive statistics, it is important to report not just a measure of central tendency (mean) but also an indication of variability. So, in parentheses after each mean below, we have included the standard deviations (*SD*).

Morning with caffeine	➔	11.8	(*SD* = 2.9)
Morning with decaf	➔	11.8	(*SD* = 2.7)
Afternoon with caffeine	➔	11.7	(*SD* = 2.8)
Afternoon with decaf	➔	8.9	(*SD* = 3.0)

So, if the word gets out about this study, the average age of Starbucks' clients might start to go up, starting around 4:00 in the afternoon. Of course, they will need to avoid the decaf.

Participant Bias

People participating in psychological research cannot be expected to respond like machines. They are humans who *know* they are in an experiment. Presumably, they have been told about the general nature of the research during the informed consent process, but in deception studies, they also know (or at least suspect) they haven't been told everything. Furthermore, even if there is no deception in a study, participants may not believe it—after all, they are in a psychology experiment and aren't psychologists always trying to psychoanalyze or manipulate people? In short, **participant bias** can occur in several ways, depending on what participants are expecting and what they believe their role should be in the study. The forms of participant bias often interact.

[4] Note that when researchers use random assignment, they assume the procedure will produce equivalent groups, especially if the groups are large enough. But, they often check to see if random assignment has done its job. In this case, they collected data from the seniors on such factors as their education level and average caffeine consumption. In Chapter 9, you will learn that this kind of information is called *demographic* data, and it is a common feature of most research in psychology. After collecting these demographic data, the experimenters determined no differences between the groups existed, which meant their random assignment procedure worked.

When behavior is affected by the knowledge that one is in an experiment and is therefore important to the study's success, the phenomenon is sometimes called the **Hawthorne effect**, after a famous series of studies of worker productivity. To understand the origins of this term, you should read Box 6.2 before continuing. You may be surprised to learn that most historians believe the Hawthorne effect has been misnamed and that the data of the original study might have been distorted for political reasons.

BOX 6.2 ORIGINS—Productivity at Western Electric

The research that led to naming the so-called Hawthorne effect took place at the Western Electric Plant in Hawthorne, Illinois, over a period of about 10 years, from 1924 to 1933. According to the traditional account, the purpose of the study was to investigate factors influencing worker productivity. Numerous experiments were completed, but the most famous series became known as the Relay Assembly Test Room study.

In the Relay Assembly experiment, six female workers were selected from a larger group in the plant. Their job was to assemble relays for the phone company. Five workers did the actual assembly, and the sixth supplied them with parts. The assembly was a time-consuming, labor-intensive, repetitive job requiring the construction of some 35 parts per relay. Western Electric produced about seven million relays a year (Gillespie, 1988), so naturally they were interested in making workers as productive as possible.

The first series of relay studies extended from May 1927 through September 1928 (Gillespie, 1988). During that time, several workplace variables were studied (and confounded with each other, actually). At various times, there were changes in the scheduling of rest periods, total hours of work, and bonuses paid for certain levels of production. The standard account has it that productivity for this small group quickly reached high levels and stayed there even when working conditions deteriorated. The example always mentioned concerned the infamous "12th test period" when workers were informed the work week would increase from 42 to 48 hours and that rest periods and free lunches would be discontinued. Virtually all textbooks describe the results somewhat like this:

> With few exceptions, no matter what changes were made—whether there were many or few rest periods, whether the workday was made longer or shorter, and so on—the women tended to produce more and more telephone relays.
>
> (Elmes, Kantowitz, & Roediger, 2006, p. 150)

Supposedly, the workers remained productive because they believed they were a special group and the focus of attention—they were part of an experiment. This is the origin of the concept called the *Hawthorne effect*, the tendency for performance to be affected because people know they are being studied in an experiment. The effect may be genuine, but whether it truly happened at Western Electric is uncertain.

A close look at what actually happened reveals interesting alternative explanations. First, although accounts of the study typically emphasize how delighted the women were to be in this special testing room, the fact is that of the five original assemblers, two had to be removed from the room for insubordination and low output. One was said to have "gone Bolshevik" (Bramel & Friend, 1981). (Remember, the Soviet Union was brand new in the 1920s, and the "red menace" was a threat to industrial America, resulting in, among other things, a fear of labor unions.) Of the two replacements, one was especially talented and enthusiastic and quickly became the group leader. She apparently was selected because she "held the record as the fastest relay-assembler in the regular department" (Gillespie, 1988, p. 122). As you might suspect, her efforts contributed substantially to the high level of productivity.

A second problem with interpreting the relay data is a simple statistical conclusion validity issue. In the famous 12th period, productivity was recorded as output per week rather than output per hour, yet workers were putting in an extra six hours per week compared to the previous test period. If the more appropriate output per hour is used, productivity actually *declined* slightly (Bramel & Friend, 1981). Also, the women were apparently angry about the change, but afraid to complain lest they be removed from the test room, thereby losing potential bonus money. Lastly, in some of the Hawthorne experiments, increased worker productivity could have been simply the result of feedback about performance, along with rewards for productivity (Parsons, 1974).

(continued)

BOX 6.2 (CONTINUED)

Historians argue that events must be understood within their entire political/economic/institutional context, and the Hawthorne studies are no exception. Painting a glossy picture of workers unaffected by specific working conditions and more concerned with being considered special ushered in the human relations movement in industry and led corporations to emphasize the humane management of employees in order to create one big happy family of labor and management. However, this picture also helps maintain power at the level of management and impede efforts at unionization (considered by managers in the 1930s to be a step toward Communism), which some historians (e.g., Bramel & Friend, 1981) believe were the true motives behind the studies completed at Western Electric.

Most research participants, in the spirit of trying to help the experimenter and contribute meaningful results, perhaps part of their Hawthorne feeling of being special, take on the role of the **good subject**, first described by Orne (1962). There are exceptions, of course, but, in general, participants tend to be cooperative, to the point of persevering through repetitive and boring tasks, all in the name of psychological science. Besides being good subjects (and maybe trying to confirm what they think is the hypothesis), research participants also wish to be perceived as competent, creative, emotionally stable, and so on. The belief that they are being evaluated in the experiment produces what Rosenberg (1969) called **evaluation apprehension**. Participants want to be evaluated positively, so they may behave as they think the ideal person should behave. This concern over how one is going to look and the desire to help the experimenter often leads to the same behavior among participants, but sometimes the desire to create a favorable impression and the desire to be a good subject conflict. For example, in a helping behavior study, astute participants might guess they are in the condition of the study designed to reduce the chances that help will be offered—the experimenter doesn't want them to help. On the other hand, altruism is a valued, even heroic, behavior in society. The pressure to be a good subject and support the hypothesis pulls the participant toward non-helping, but evaluation apprehension makes the individual want to help. At least one study has suggested that when participants are faced with the option of confirming the hypothesis and being evaluated positively, the latter is the more powerful motivator (Rosnow, Goodstadt, Suls, & Gitter, 1973).

Furthermore, if participants can figure out the hypothesis, they may try to behave in a way that confirms it. Orne (1962) used the term **demand characteristics** to refer to those aspects of the study that reveal the hypotheses being tested. If these features are too obvious to participants, participants may no longer act naturally; instead, they may behave the way they think they are supposed to behave, making it difficult to interpret the results. Did participants behave as they normally would or did they come to understand the hypothesis and behave so as to make it come true or even to defy it? The presence of demand characteristics can severely reduce a study's internal validity. The possibility that demand characteristics are operating can affect the choice of between- or within-subject designs. Participants serving in all of the conditions of a study have a greater opportunity to figure out the hypothesis. Hence, demand characteristics are potentially more troublesome in within-subject designs than in between-subjects designs. For both types, demand characteristics are especially devastating if they affect some conditions but not others, thereby introducing a confound.

Demand characteristics can operate in many subtle ways. Here is an example of a study showing how they might operate in a research area that has become an important one for human well-being.

Research Example 10—Demand Characteristics

With obesity a major health risk, psychologists for some time have been interested in studying the factors that affect eating behavior. The result has been a long list of reasons why people often eat even if they are not especially hungry, ranging from emotional factors such as depression to situational cues such as the visual attractiveness of food. A common procedure is to give subjects

the opportunity to evaluate food in some form of taste test, while giving them the opportunity to eat as much of the sample food as they would like. The researcher will be more interesting in measuring how much is eaten rather than the taste evaluations. Robinson, Kersbergen, Brunstrom, and Field (2014) wondered if demand characteristics could be operating in these eating behavior studies. In particular, they were concerned that if subjects detected that "amount of food eaten" was the dependent measure of interest, they might eat less than they normally would eat.

After conducting an online survey showing that people reported that they would be likely to reduce their eating behavior if they suspected their eating was being monitored, Robinson et al. (2014) designed a study to see if cues about the experimenter monitoring food consumption would in fact reduce eating behavior. Subjects were randomly assigned to three groups in a study using a standard taste-test procedure. Thus, all participants were told that they would be evaluating the taste of a batch of cookies, that "they were free to eat as many cookies as they liked, and that any remaining food would be thrown away" (p. 22). This is all that subjects in a control condition were told; those assigned to a "monitored" condition were also told that the researcher would be recording how many cookies were eaten; those in an "unmonitored" condition were told to dispose of remaining cookies in a waste bin after they finished their taste ratings. This third condition was included in an attempt "to convince the participants that their food consumption was not being monitored" (p. 22).

The results clearly indicated that eating behavior was affected by knowledge of food monitoring—subjects ate significantly less when they knew the number of cookies they were eating would be counted. Here are the means and standard deviations (the dependent variable was the number of grams of cookie eaten):

Control condition	$M = 45.8$	$SD = 21.9$
Monitored condition	$M = 29.3$	$SD = 12.4$
Unmonitored condition	$M = 47.7$	$SD = 25.3$

As a *manipulation check*, Robinson et al. (2014) asked subjects at the end of the study to indicate on a 5-point scale whether they believed their eating behavior was being monitored (regardless of what their instructions had been). These data were

Control condition	$M = 3.7$	$SD = 1.1$
Monitored condition	$M = 4.0$	$SD = 0.8$
Unmonitored condition	$M = 3.3$	$SD = 0.9$

These means were not significantly different from each other.

Taken together, these results indicate that if subjects in eating behavior studies suspect the amount of their food consumption is being measured, they will significantly reduce their eating behavior. This is the classic instance of a demand characteristic – knowledge of a researcher's purpose affecting participant behavior. One particular danger that Robinson et al. (2014) pointed out was that this demand characteristic could result in some eating behavior studies finding no significant results because of a *floor effect*. That is, in a study designed to show that some factor reduces eating, the demand characteristic might reduce eating to low levels in all the conditions of the study, making it impossible to detect differences among conditions.

Controlling for Participant Bias

The primary strategy for controlling participant bias is to reduce demand characteristics to the minimum. One way of accomplishing this, of course, is through deception. As we've seen in Chapter 2, the primary purpose of deception is to induce participants to behave more naturally than they otherwise might. A second strategy, normally found in drug studies, is to use a placebo

control group (elaborated in Chapter 7). This procedure allows for comparison of those actually getting some treatment (e.g., a drug) and those who *think* they are getting the treatment but aren't. If the people in both groups behave identically, the effects can be attributed to participant expectations of the treatment's effects. You have probably already recognized that the caffeine study you read about (Research Example 9) used this form of logic.

A second way to check for the presence of demand characteristics is to do a *manipulation check***.** This can be accomplished during debriefing by asking participants in a deception study to indicate what they believe the true hypothesis to be (the "good subject" might feign ignorance, though). You recall that this was the strategy used in Research Example 9 by asking participants to guess whether they had been given caffeine in their coffee or not. Manipulation checks can also be performed during an experiment. Sometimes, a random subset of participants in each condition will be stopped in the middle of a procedure and asked about the clarity of the instructions, what they think is going on, and so on. In the study on violent video games, desensitization, and helping that you read about in Chapter 3 (Research Example 1), the manipulation check was built directly into the procedure in the form of a survey asking subjects to rate the violence level of the video games. Manipulation checks are also used to see if some procedure is producing the effect it is supposed to produce. For example, if a procedure is supposed to make people feel anxious (e.g., telling participants to expect shock), a sample of participants might be stopped in the middle of the study and assessed for level of anxiety.

A final way of avoiding demand characteristics is to conduct field research. If participants are unaware they are in a study, they are unlikely to spend any time thinking about research hypotheses and reacting to demand characteristics. Of course, field studies have problems of their own, as you recall from the discussion of informed consent in Chapter 2.

Although we stated earlier that most research participants play the role of "good subjects," this is not uniformly true, and differences exist between those who truly volunteer and are interested in the experiment and those who are more reluctant and less interested. For instance, true volunteers tend to be slightly more intelligent and have a higher need for social approval (Adair, 1973). Differences between volunteers and non-volunteers can be a problem when college students are asked to serve as participants as part of a course requirement; some students are more enthusiastic volunteers than others. Furthermore, a semester effect can operate. The true volunteers, those really interested in participating, sign up earlier in the semester than the reluctant volunteers. Therefore, if you ran a study with two groups, and one group was tested in the first half of the semester and the other group in the second half, the differences found could be due to the independent variable, but they also could be due to differences between the true volunteers who signed up first and the reluctant volunteers who waited as long as they could to sign up. Can you think of a way to control for this problem? If "blocked random assignment" occurs to you, and you say to yourself "This will distribute the conditions of the study equally throughout the duration of the semester," then you've accomplished something in this chapter. Well done!

Self Test 6.3

1. Unlike most longitudinal studies, Terman's study of gifted children did not experience which control problem?
2. Why does a double blind procedure control for experimenter bias?
3. How can a demand characteristic influence the outcome of a study?

To close this chapter, read Box 6.3, which concerns the ethical obligations of participants in psychological research. The list of responsibilities you'll find there is based on the assumption that research should be a collaborative effort between experimenters and participants. We've seen that experimenters must follow the APA ethics code. In Box 6.3 you'll learn that participants have responsibilities too.

In the last two chapters, you have learned about the essential features of experimental research and some of the control problems faced by those who wish to do research in psychology. We've now completed the necessary groundwork for introducing the various experimental designs used to test the effects of independent variables. So, let the designs begin!

BOX 6.3 ETHICS—Research Participants Have Responsibilities Too

The APA ethics code spells out the responsibilities researchers have to those who participate in their experiments. Participants have a right to expect that the guidelines will be followed, and the process for registering complaints should be a clear if they are not. But what about the participants? What are their obligations?

An article by Korn in the journal *Teaching of Psychology* (1988) outlines the basic rights that college students have when they participate in research, but it also lists the responsibilities of those who volunteer. These include:

- Be responsible about scheduling by showing up on time for appointments with researchers.
- Be cooperative and acting professionally by giving their best and most honest effort.
- Listen carefully to the experimenter during the informed consent and instructions phases and asking questions if they are not sure what to do.
- Respect any request by the researcher to avoid discussing the research with others until all the data are collected.
- Be active during the debriefing process by helping the researcher understand the phenomenon being studied.

The assumption underlying this list is that research should be a collaborative effort between researchers and participants. Korn's (1988) suggestion that subjects take a more assertive role in making research collaborative is a welcome one. This assertiveness, however, must be accompanied by enlightened experimentation that values and probes for the insights participants have about what might be going on in a study. An experimenter who simply runs a subject and records the data is ignoring valuable information.

CHAPTER SUMMARY

Between-Subjects Designs

In between-subjects designs, individuals participate in just one of the experiment's conditions; hence, each condition in the study involves a different group of participants. Such a design is usually necessary when subject variables (e.g., gender) are being studied or when being in one condition of the experiment changes participants in ways that make it impossible for them to participate in another condition. With between-subjects designs, the main difficulty is creating groups that are essentially equivalent on all factors except for the independent variable.

Creating Equivalent Groups

The preferred method of creating equivalent groups in between-subjects designs is random assignment. Random assignment has the effect of spreading unforeseen confounding factors evenly throughout the different groups, thereby eliminating their damaging influence. The chance of random assignment working to produce equivalent groups increases as the number of participants per group increases. If few participants are available, if some factor (e.g., intelligence) correlates highly with the dependent variable, and if that factor can be assessed without difficulty before the experiment

begins, then equivalent groups can be formed by using a matching procedure. Subjects are matched on some matching variable and then randomly assigned to groups.

Within-Subjects Designs

When each individual participates in all of the study's conditions, the study is using a within-subjects (or repeated-measures) design. For these designs, participating in one condition might affect how participants behave in other conditions—that is, sequence or order effects can occur, both of which can produce confounded results if not controlled. Order effects include progressive effects (they gradually accumulate, as in fatigue or boredom) and carryover effects (one sequence of conditions might produce effects different from another sequence). When substantial carryover effects are suspected, researchers usually switch to a between-subjects design.

Controlling Order Effects

Order effects are controlled by counterbalancing procedures that ensure the conditions are tested in more than one sequence. When participants serve in every condition of the study just once, complete (all possible orders of conditions used) or partial (a sample of different orders or a Latin square) counterbalancing is used. When participants serve in every condition more than once, reverse counterbalancing or blocked randomization can be used.

Methodological Control in Developmental Research

In developmental psychology, the major independent variable is age, a subject variable. If age is studied between subjects, the design is cross sectional. It has the advantage of efficiency, but cohort effects can occur, a special form of the problem of nonequivalent groups. If age is a within-subjects variable, the design is longitudinal and attrition can be a problem. The two strategies can be combined in a cohort sequential design: selecting new cohorts every few years and testing each cohort longitudinally.

Controlling for the Effects of Bias

The results of research in psychology can be biased by experimenter expectancy effects. These can lead the experimenter to treat participants in different conditions in various ways, making the results difficult to interpret. Such effects can be reduced by automating the procedures and/or using double blind control procedures. Participant bias also occurs. Participants might behave in unusual ways simply because they know they are in an experiment or they might confirm the researcher's hypothesis if demand characteristics suggest to them the true purpose of a study.Demand characteristics are usually controlled through varying degrees of deception; the extent of participant bias can be evaluated by using a manipulation check.

CHAPTER REVIEW QUESTIONS

1. Under what circumstances would a between-subjects design be preferred over a within-subjects design?
2. Under what circumstances would a within-subjects design be preferred over a between-subjects design?
3. How does random selection differ from random assignment, and what is the purpose of each?
4. As a means of creating equivalent groups, when is matching more likely to be used than random assignment?
5. Distinguish between progressive effects and carryover effects, and explain why counterbalancing might be more successful with the former than the latter.
6. In a taste test, Joan is asked to evaluate four dry white wines for taste: wines A, B, C, and D. In what sequence would they be tasted if (a) reverse counterbalancing or (b) blocked randomization were being used? How many orders would be required if the researcher used complete counterbalancing?
7. What are the defining features of a Latin square, and when is one likely to be used?
8. What specific control problems exist in developmental psychology with (a) cross-sectional studies and (b) longitudinal studies?
9. What is a cohort sequential design, and why is it an improvement on cross-sectional and longitudinal designs?
10. Describe an example of a study that illustrates experimenter bias. How might such bias be controlled?
11. What are demand characteristics and how might they be controlled?
12. What is a Hawthorne effect? What is the origin of the term?

APPLICATIONS EXERCISES

Exercise 6.1. Between-Subject or Within-Subject?

Think of a study that might test each of the following hypotheses. For each, indicate whether you think the independent variable should be a between- or a within-subjects variable or whether either approach would be reasonable. Explain your decision in each case.

1. A neuroscientist hypothesizes that damage to the primary visual cortex is permanent in older animals.
2. A sensory psychologist predicts that it is easier to distinguish slightly different shades of gray under daylight than under fluorescent light.
3. A clinical psychologist thinks that phobias can be cured by repeatedly exposing the person to the feared object and not allowing the person to escape until the person realizes the object really is harmless.
4. A developmental psychologist predicts cultural differences in moral development.
5. A social psychologist believes people will solve problems more creatively when in groups than when alone.
6. A cognitive psychologist hypothesizes that giving subjects repeated tests of verbal information will lead to greater retention than asking subjects to study the verbal information repeatedly.
7. A clinician hypothesizes that people with obsessive-compulsive disorder will be easier to hypnotize than people with a phobic disorder.
8. An industrial psychologist predicts that worker productivity will increase if the company introduces flextime scheduling (i.e., work eight hours, but start and end at different times).

Exercise 6.2. Fixing Confounds

Return to Exercise 5.2 in Chapter 5 and fix the confounds in those studies by designing a well-controlled study for each scenario. For each study, be sure to explain how you would use the methodological controls you learned about in Chapter 6.

Exercise 6.3. Random Assignment and Matching

A researcher investigates the effectiveness of an experimental weight-loss program. Sixteen volunteers will participate, half assigned to the experimental program and half placed in a control group. In a study such as this, it would be good if the average weights of the subjects in the two groups were approximately equal at the start of the experiment. Here are the weights, in pounds, for the 16 subjects before the study begins.

168	210	182	238	198	175	205	215
186	178	185	191	221	226	188	184

First, use a matching procedure as the method to form the two groups (experimental and control) and then calculate the average weight per group. Second, assign participants to the groups again, this time using random assignment (cut out 20 small pieces of paper, write one of the weights on each, and then draw them out of a hat to form the two groups). Again, calculate the average weight per group after the random assignment has occurred. Compare your results to those of the rest of the class. Are the average weights for the groups closer to each other with matching or with random assignment? In a situation such as this, what do you conclude about the relative merits of matching and random assignment?

Answers to Self Tests

✓6.1

1. A minimum of two groups of subjects is tested in the study, one group for each level of the IV; the problem of equivalent groups.
2. As in the Barbara Helm study, it is sometimes essential that subjects not be aware of what occurs in other conditions of the study.
3. Sal must have a reason to expect verbal fluency to correlate with his dependent variable; he must also have a good way to measure verbal fluency.

✓6.2

1. Each subject participates in each level of the IV; order effects
2. With six levels of the IV, complete counterbalancing requires a minimum of 720 subjects (6 x 5 x 4 x 3 x 2 x 1), which could be impractical.
3. Reverse counterbalancing or block randomization.

✓6.3

1. Attrition.
2. If the experimenter does not know which subjects are in each of the groups in the study, the experimenter cannot behave in a biased fashion.
3. If subjects know what is expected of them, they might be "good subjects" and not behave naturally.