Chapter 6

Control Problems in Experimental Research

Preview & Chapter Objectives

In Chapter 5 you learned the essentials of the experimental method—manipulating an independent variable, controlling everything else, and measuring the dependent variable. In this chapter we will begin by examining two general types of experimental design, one in which different groups of participants contribute data for different levels of the independent variable (between-subjects design) and one in which the same participants contribute data to all the levels of the independent variable (within-subjects design). As you are about to learn, there are special advantages associated with each approach, but there are also problems that have to be carefully controlled—the problem of equivalent groups for between-subjects designs, and problems of sequence for within-subjects designs. The last third of the chapter addresses the issue of bias and the ways of controlling it. When you finish this chapter, you should be able to:

- Discriminate between-subjects designs from within-subjects designs.
- Understand how random assignment can solve the equivalent groups problem in between-subjects designs.
- Understand when matching should be used instead of random assignment when attempting to create equivalent groups.
- Distinguish between progressive and carryover effects in within-subjects designs, and understand why counterbalancing normally 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 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.

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 he put it, "[w]hether 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." (Woodworth, 1938, p. 3). Some of these difficulties we've 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'll look at some other aspects of maintaining control: the problem of creating equivalent groups in experiments involving separate groups of participants, the problem of sequence effects in experiments in which participants 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 condition A with condition 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 levels A and B will be a contrast *between* two different groups of individuals. On the other hand, if each participant receives both levels A *and* B, you could say that both levels exist *within* each individual; 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. Unless the researcher could round up some multiple personalities, introverted in one personality and extroverted in another, there is no alternative but to compare two different groups. One of the few times a subject variable won't be a between-subject variable is when behaviors occurring at two different ages are being compared, and the same persons are studied at two different times in their lives. Another possibility is when marital status is the subject variable, and the same people are studied before and after a marriage or a divorce. Most of the time, however, using a subject variable means that a between-subjects design will be used.

Using a between-subjects design is unavoidable in some studies that use certain manipulated independent variables. That is, it is sometimes the case that 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 of it. There were two separate between-subjects manipulated independent variables. One 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" (Sigall & Ostrove, 1975, 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 an unattractive Barbara (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 (2.8 years) than unattractive (5.2) or control (5.1) Barbara. However, the opposite happened when the crime was swindling. Apparently thinking that Barbara was using her good looks to commit the crime, participants gave attractive Barbara a harsher sentence (5.5 years) than they gave the unattractive (4.4) or control (4.4) woman.

You can see why it was necessary to run this study with between-subjects 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 participant enters the study fresh, and naive with respect to the procedures to be tested, the prime disadvantage is that large numbers of people may need to be recruited, tested, and debriefed. Hence, the researcher invests a great deal of energy in this type of design. My doctoral dissertation on memory involved five different experiments requiring between-subjects factors; more than 600 students trudged in and out of my lab before the project was finished!

Another disadvantage of between-subjects designs is that differences between the conditions could be due to the independent variables, but they might also be due to differences between the two groups. To deal with this potential confound, deliberate steps must be taken to create what are called **equivalent groups.** These groups are equal to each other in every important way except for the levels of the independent variable. The number of equivalent groups in a between-subjects study corresponds exactly to the number of different conditions in the study, with one group of participants tested in each condition.

The Problem of Creating Equivalent Groups

There are two common techniques for creating equivalent groups in a between-subjects experiment. The ideal approach is to use random assignment. A second strategy is to use matching.

Random Assignment

First, be sure you understand that random assignment and random selection are not the same. Random selection, to be described in Chapter 12 (pp. xx), is a procedure for getting volunteers to come into your study. As you will learn, it is a process designed to produce a sample of individuals who reflect the broader population, and it is a common strategy in research using surveys. Random assignment is a method for placing participants, once selected for a study, into the different groups. When **random assignment** is used, every person volunteering for the study has an equal chance of being placed in any of the groups 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're comparing two presentation rates in a simple memory study. Further suppose that anxious participants won't do as well on your memory task as nonanxious participants, but you as the researcher are unaware of that fact. 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 for the 4-second group. 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, I've placed an "A" or an "R" in parentheses as a way of telling you which participants are anxious and which are relaxed. Data for the anxious people are shaded.

If you look carefully at these data, you'll see that the three anxious participants in each group did worse than their five relaxed peers. Because there are an equal number of anxious participants in each group, 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—an average of 15 words for the 2-second group and 19 for the 4-second group.

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

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 different groups. This is especially true when large numbers of individuals are being assigned to each group. In fact, the greater the number of participants involved, the greater the chance that random assignment will work to create equivalent groups of them. If groups are equivalent and if all else is adequately controlled, then you are in that enviable position of being able to say that your independent variable was responsible if you find differences between your groups.

You might think the actual process of random assignment would be fairly simple—just use a table of random numbers to assign each arriving participant to a group or, in the case of a two-group study, flip a coin. Unfortunately, however, 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?

To complete a random assignment of participants to conditions in a way that guarantees an equal number of participants per group, a researcher can use **block randomization,** 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 in actual practice researchers typically rely on a simple computer program to generate a sequence of conditions meeting the requirements of block randomization—you can find one at http://www.randomizer.org/.

Matching

When only a small number of subjects are available for your experiment, random assignment can sometimes fail to create equivalent groups. The following example shows you how this might happen. Let's take the same study of the effect of presentation rate on memory, used earlier, and assume that 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:[1]

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 (17 = 17) and would wonder what happened. After all, participants were randomly assigned, and the researcher's prediction about better recall for a slower presentation rate certainly makes sense. So what went wrong?

What happened was that random assignment 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 wiped out 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 the anxiety effect. 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; this study just failed to find the effect). To repeat what was mentioned earlier, the chance of random assignment working to create equivalent groups increases as sample size increases.

To deal with the problem of equivalent groups in a situation such as this, a matching procedure could be used. In **matching,** participants are grouped together on some trait such as anxiety level, and then distributed randomly to the different

[1] This same pattern of results could occur if an experimenter failed to randomly assign and naively tested the first eight people to sign up in the 2-second rate group and the next eight people in the other group. It is conceivable that the more anxious students would delay volunteering to participate, increasing the chances of their being placed in the 4-second group.

groups in the experiment. In the memory study, "anxiety level" would be called a **matching variable.** Individuals in the memory experiment would be given some reliable and valid measure of anxiety, those with similar scores would be paired together, and one person in each pair would be randomly placed in the group getting the 2-second rate and the other would be put into the group with the 4-second rate. As an illustration of exactly how to accomplish matching in a two-group experiment, you should work through the example in Table 6.1.

Matching sometimes is used when the number (*N*) of participants is small, and random assignment is therefore risky and might yield nonequivalent groups. In order to undertake matching, however, two important conditions must be met. First, you must have good reason to believe that the matching variable will have a predictable effect on the outcome of the study. That is, you must be confident that the matching variable is correlated with the dependent variable. This was the case in our hypothetical memory study—anxiety clearly reduced recall. When there is a high correlation between the matching variable and the dependent variable, the statistical techniques for evaluating matched-groups designs are sensitive to differences between the groups. On the other hand, if matching is done when there is a low correlation between the matching variable and the dependent variable, the chances of finding a true difference between the groups decline. So it is important to be careful when picking matching variables.

A second important condition for matching is that there must be some 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 participants (e.g., Grade Point Average scores or IQ from school records), or by matching on the dependent variable itself. That is, in a memory study, participants could be given an initial memory test, then matched on their performance, and then assigned to 2-second and 4-second groups. Their preexisting memory ability would thereby be under control and the differences in performance could be attributed to the presentation rate.

One practical difficulty with matching concerns the number of matching variables to use. In a memory study, should I match the groups for anxiety level? What about intelligence level? What about education level? You can see that some judgment is required here, for matching is difficult to accomplish with more than one matching variable, and often results in having to eliminate participants because close matches sometimes cannot be made. The problem of deciding on and measuring matching variables is one reason why 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 anxiety levels, intelligence, or education level. They simply make the groups large enough and assume that random assignment will distribute these potentially confounding factors evenly throughout the conditions of the study.

TABLE 6.1 *How to Use a Matching Procedure*

In a study on problem solving requiring two different groups, a researcher is concerned that a participant's academic skills may correlate highly with performance on the problems to be used in the experiment. The participants are college students, so the researcher decides to match the two groups on grade point average (GPA). That is, deliberate steps will be taken to insure that the two groups are equivalent to each other in academic ability, as reflected in their average GPAs. Here's how it is done:

Step 1. Get a score for each person on the matching variable. That's easy in this case because it simply means retrieving GPA data from the Registrar (with the students' consent of course). In other cases of matching, the matching variable must be determined by pretesting participants on the variable; this can mean bringing participants to the lab twice, which can be inconvenient (another reason why researchers like random assignment).

Suppose there will be 10 volunteers (Ss) in the study, 5 per group. Here are their GPAs:

S1:	3.24	S6:	2.45
S2:	3.91	S7:	3.85
S3:	2.71	S8:	3.12
S4:	2.05	S9:	2.91
S5:	2.62	S10:	2.21

Step 2. Arrange the GPAs in ascending order:

S4:	2.05	S9:	2.91
S10:	2.21	S8:	3.12
S6:	2.45	S1:	3.24
S5:	2.62	S7:	3.85
S3:	2.71	S2:	3.91

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

Pair 1:	2.05 and 2.21
Pair 2:	2.45 and 2.62
Pair 3:	2.71 and 2.91
Pair 4:	3.12 and 3.24
Pair 5:	3.85 and 3.91

Step 4. For each pair, randomly assign one participant to Group 1 and the other to Group 2. Here's one possible outcome:

	Group 1	*Group 2*
	2.05	2.21
	2.62	2.45
	2.91	2.71
	3.12	3.24
	3.85	3.91
mean GPA:	**2.91**	**2.90**

Now the study can proceed with some assurance that the two groups will be equivalent to each other (2.91 is virtually the same as 2.90) in terms of academic ability.

Note. If more than two groups are being tested,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 participants in each cluster are randomly assigned to the multiple groups.

✓ 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. Sal wishes to see if the type of 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 the 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, you will sometimes see this procedure 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 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 such areas as physiological psychology and sensation and perception, comparisons often are made between conditions that require just a brief amount of 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 press a key that changes the length of one of the lines. Participants are told to adjust the line until both lines seem 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 someone 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

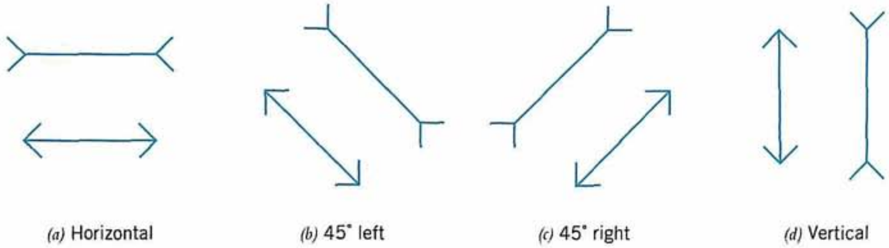

FIGURE 6.1 Set of four Müller-Lyer illusions: horizontal, 45° left, 45° right, vertical.

duplicate each level several times). And unlike the attractive/unattractive Barbara Helm study, serving in one condition would not make it impossible to serve in another.

One of psychology's oldest areas of research is in psychophysics, the study of sensory thresholds (e.g., a modern application is a hearing test). In a typical psychophysics study, subjects are asked to judge whether or not they can detect some stimulus or whether two stimuli are equal or different. Each situation requires a large number of trials and comparisons to be made within the same individual. Hence, psychophysics studies typically use just a few participants and measure them repeatedly. Research Example 5, which you will soon encounter, uses this strategy.

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) are just two examples. Of course, there are times when, even with a limited population, the design may require a between-subjects manipulation. Evaluating the effects of a new form of therapy for those suffering from a rare form of psychopathology requires comparing those in therapy with others in a control group not being treated.

Besides convenience, another advantage of within-subjects designs is that they eliminate the equivalent groups problem that occurs with between-subjects designs. Recall from Chapter 4 that an inferential statistical analysis comparing two groups examines the variability between experimental conditions with the variability within each condition. Variability between conditions could be due to (a) the independent variable, (b) other systematic variance resulting from confounding, and/or (c) nonsystematic error variance. Even with random assignment, a significant portion of the error variance in a between-subjects design results 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 concrete 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 warming up, each golfer hits one ball or the other. Here are the results:

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**

There are several things to note here. First, there is some variability within each group, as reflected in the standard deviation for each group. This is error variance due to individual differences within each group and to other random factors.

Second, there is apparently an overall difference between the groups. The pros in the second group hit their ball farther than the pros in the first group. Why? Three possibilities:

a. Chance; perhaps this is not a statistically significant difference, and even if it is, there's a 5% chance that it is a Type I error if the null hypothesis is actually true.

b. The golf ball; perhaps the brand of golf ball hit by the second group simply goes farther (this, of course, is the research hypothesis).

c. Individual differences; maybe the golfers in the second group are stronger or more skilled than those in the first group.

The chances that the third 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 that 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 back to the individual differences between the two groups. This problem simply does not occur in a within-subjects design. Suppose you repeated the study but used just the first five golfers, and each pro hits 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, explanation 3 can be eliminated for the second set. In the first set, the difference in the first row between the 255 and the 269 could be due to chance, the difference between the balls, or individual differences between pro 1 and pro 6. In the second set, there is no second group of golfers, so the third possibility is gone. Thus, in a within-subjects design, individual differences are eliminated from the estimate of the amount of 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 will be the case for a between-subjects design.

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 (a) chance factors and/or (b) 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 20 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 in some way so that 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. Also, some particular sequences might produce effects that are different from those of other sequences, what could be called a **carryover effect**. Thus, in a study with two basic conditions, experiencing condition A before condition B might affect the person much differently than experiencing B before A. For example, suppose you were studying the effects of noise on a problem-solving task using a within-subjects design. Let's say that participants will be trying to solve anagram problems (rearrange letters to form words) under some time pressure. In condition A, 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 B, the same total amount of noise occurs; however, it is not randomly presented but instead occurs in predictable patterns. If you put the people in condition A first (unpredictable noise), and then in B (predictable noise), they will probably do poorly in A (most people do). This poor performance might discourage them and carry over to condition B. They should do better in B, 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 B first, with 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. When they then encounter condition A, they might do better than you would ordinarily expect. Thus, performance in condition A might be much worse in the sequence A–B than in the sequence B–A, and a similar problem would occur for condition B. In short, the sequence in which the conditions are presented, independently of any practice or fatigue effects, might influence the study's outcome. In studies where carryover effects might be suspected, researchers often switch to a between-subjects design. Indeed, studies comparing predictable and unpredictable noise typically put people in two different groups.

The Problem of Controlling Sequence Effects

The normal way to control sequence effects in a within-subjects design is to use more than one sequence, a strategy known as **counterbalancing**. As I will elaborate later, 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 will be tested in each of the conditions but tested 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 six different 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 that the pair of players of 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 that 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.

You'll recognize the design of the Reynolds study as including a within-subjects variable. Each of the 15 participants examined all six games. Also, you can see that 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 any sequence effects that might be present. He certainly didn't want all 15 participants to see the six 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 sequence problem is to use **complete counterbalancing**. This means that 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 sequences 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

The problem with complete counterbalancing is that as the number of levels of the independent variable increases, the possible sequences that will be needed increase exponentially. There are 6 sequences needed for three conditions, but simply adding a fourth condition creates a need for 24 sequences ($4 \times 3 \times 2 \times 1$). As you can guess, complete counterbalancing was not possible in Reynolds' study unless he recruited many more than 15 chess players. In fact, with six different games (i.e., conditions), he would need to find 6! or 720 players to cover all of the possible sequences. Clearly, Reynolds used a different strategy.

Partial Counterbalancing

Whenever a subset of the total number of sequences is used, the result is called **partial counterbalancing**. This was Reynolds's solution; he simply took a random sample of the 720 possible sequences by ensuring that "the order of presentation [was] randomized for each subject" (Reynolds, 1992, p. 411). Sampling from the population of sequences is a common strategy whenever there are fewer participants available than possible sequences or when there are a fairly large number of conditions.

Reynolds sampled from the total number of sequences, 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 that each letter appears only once in each row and each column (Kirk, 1968). The Latin square strategy is more sophisticated than simply 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. Work through Table 6.2 to see how to construct the following 6 × 6 Latin square. 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

I've boldfaced condition **A** (chess game A) 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 one time. From the top row to the bottom, (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. To use the 6 × 6 Latin square, one randomly assigns each of the six conditions of the experiment (six different chess games for Reynolds) to one of the six letters, A through F.

When using Latin squares, it is necessary for the number of participants to be equal to or be a multiple of the number of rows in the square. The fact the Reynolds had 15 participants in his study tells you that 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 six rows of the square (3 × 6 = 18).

Testing More Than Once per Condition

In the Reynolds 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 × 4 Latin square, they will seldom be asked to study and recall any particular list a second time unless the researcher is specifically interested in the effects of repeated

TABLE 6.2 *Building a Balanced 6 × 6 Latin Square*

In a balanced Latin square, every condition of the study occurs equally often in every sequential position, and every condition precedes and follows every other condition exactly once. Here's how to build a 6×6 square.

Step 1. Build the first row. It is fixed according to this general rule:

A B "X" C "X – 1" D "X – 2" E, "X – 3," F, etc.

where A refers to the first condition of the study and "X" refers to the letter symbolizing the final condition of the experiment. To build the 6 × 6 square, this first row would substitute:

X = the sixth letter of the alphabet → F
X – 1 = the fifth letter → E

Therefore, the first row would be

A B **F** (subbing for "X") C E (subbing for "X – 1") D

Step 2. Build the second row. Directly below each letter of row 1, place in row 2 the letter that is next in the alphabet. The only exception is the F. Under that letter, return to the first of the six letters and place the letter A. Thus:

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

Step 3. Build the remaining four rows following the step 2 rule. Thus, the final 6 × 6 square is:

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

Step 4. Take the six conditions of the study and randomly assign them to the letters. A through F to determine the actual sequence of conditions for each row. Assign an equal number of participants to each row.

Note. This procedure works whenever there is an even number of conditions. If the number of conditions is odd, two squares will be needed—one created using the above procedure, and a second an exact reversal of the square created with the above procedure. For more details, see Winer, Brown, and Michaels (1994).

trials on memory. However, in many studies it is reasonable, even necessary, for participants to experience each condition more than one time. This often happens in research in sensation and perception, for instance. A look back at Figure 6.1 provides an example.

Suppose you were conducting a study in which you wanted to see if participants would be more affected by the illusion when it was presented vertically than when shown horizontally or at a 45° angle. Four conditions of the study are assigned to the letters A–D:

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 have to make adjustments to the lengths of the parallel lines until they perceive that the lines are equal. The four conditions could be presented 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 desires to have the participant perform the task more than twice per condition, and this is common in perception research, this sequence 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, participants would see the illusions in this sequence:

A–B–C–D—D–C–B–A—A–B–C–D—D–C–B–A—A–B–C–D—D–C–B–A

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 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 (pp. 239).[2]

Block Randomization

A second way to present a sequence of conditions when each condition is presented more than once is to use **block randomization,** the same procedure outlined earlier in the context of how to assign participants randomly to groups in a between-subjects experiment. The basic rule is that every condition occurs once before any condition is repeated a second time. Within each block, the order of conditions is randomized.

[2] 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 different 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).

This strategy eliminates the possibility that participants can predict what is coming next, a problem that can occur with reverse counterbalancing.

To use 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 auditory perception study by Carello, Anderson, and Kunkler-Peck (1998).

Research Example 5—Counterbalancing with Block Randomization

Our ability to localize sound has been known for a long time—under normal circumstances, we are quite adept at identifying the location from which a sound originates. What interested Carello and her research team was whether people could identify something about the physical size of an object simply by hearing it drop on the floor. She devised the apparatus pictured in Figure 6.2 to examine the question. Participants heard a wooden dowel hit the floor, and then tried to judge its length. They made their response by adjusting the distance between the edge of the desk they were sitting at and a movable vertical surface during a "trial," which was defined as having the same dowel dropped five times in a row from a given height. During the five drops, participants were encouraged to move the wall back and forth until they were comfortable with their decision about the dowel's size. In the first of two experiments, the within-subjects independent variable was the length of the dowel, and there were seven levels (30, 45, 60, 75, 90, 105, and 120 cm). Each participant

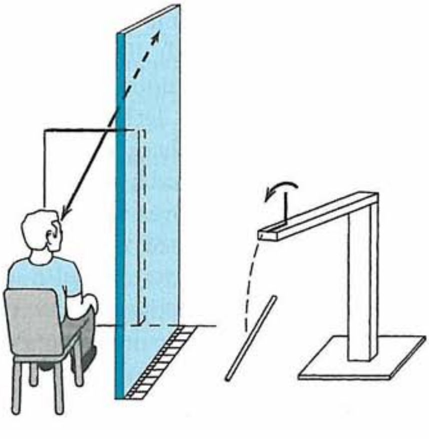

FIGURE 6.2 The experimental setup for Carello, Anderson, & Kunkler-Peck (1998). After hearing a rod drop, participants adjusted the distance between the edge of their desk and the vertical surface facing them to match what they perceived to be the length of the rod.

✓ 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 6 levels, each tested just once per subject, why are you more likely to use partial counterbalancing instead of 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?

Control Problems 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. There are some additional considerations for researchers in developmental psychology, where two specific varieties 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 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 different groups of children. A **longitudinal study,** on the other hand, studies a single group over a period of time; it takes a within-subjects or repeated-measures approach. 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 5.

The obvious advantage of the cross-sectional approach to the experiment on language is time; such a study might take a month to complete. If done as a longitudinal study, it would take 3 years. However, a potentially serious difficulty with some cross-sectional studies is a special form of the problem of nonequivalent groups and involves what are known as **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 just simply 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 30, 50, and 70? You might indeed find a decline with age, but does it mean that intelligence gradually decreases with age, or might the differences relate to the very different life histories of the three groups? For example, the 70-year-olds went to school during the Great Depression, the 50-year-olds were educated during the post–World War II boom, and the 30-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 very different pattern (Schaie, 1988). For example, verbal abilities show very little decline, especially if the person remains verbally active (moral: use it or lose it).

While cohort effects can plague cross-sectional studies, longitudinal studies also have problems, most notably with attrition (refer back to Chapter 5, p. 189). If a large number of participants drop out of the study, the group completing it may be very 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. There are also potential ethical problems in longitudinal studies. 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, some researchers use a strategy that combines cross-sectional with longitudinal studies, a design referred to as a **cohort sequential design**. In such a study, a group of subjects will be selected and retested every few years, and then additional cohorts will be selected every few years and also retested over time. 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 five 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 1960 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 #	1960	1965	1970	1975	1980	1985	1990
1	55	60	65				
2		55	60	65			
3			55	60	65		
4				55	60	65	
5					55	60	65

So in 1960, you have a group of 55-year-olds that you test. Then in 1965, these same people (now 60 years old) would be retested, along with a new group of 55-year-olds. By year 3, you have cohorts for all three age groups. As you can see, combining the data in each of the diagonals would give you an overall comparison between those aged 55, 60, and 65. Comparing the data in the rows enables a comparison of overall differences between cohorts. In actual practice, these deigns are more complicated, because researchers will typically start the first year of the study with a 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. It began in 1956, designed to examine age-related changes in various mental abilities. The initial cohort had 500 people in it, ranging in age from their early 20s to their late 60s (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. In all, about 6,000 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 twentieth century have had the advantages of better education, better nutrition, and so on.

The length of Schaie's Seattle project is impressive, but the world's 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. His doctoral dissertation, supervised by Edmund Sanford at Clark University in 1905, was his first serious investigation of giftedness; in it, he compared 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 took the intelligence test created by Alfred Binet of France and transformed it into the popular Stanford-Binet IQ test. Second, he began a longitudinal study of gifted children that continued long after he himself died.

Terman was motivated by the belief, shared by most mental testers of his day, that America 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) talent and nurture it.

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% of the population. Each child 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 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. ix). 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).

There are three points worth making about this mega-longitudinal study. First, Terman's work shattered the stereotype of the gifted child as someone who was brilliant but socially retarded 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. 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 that 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 his special people. After all, the group represented the type of person Terman believed held the key to America's future.

Problems with Biasing

Because humans are always the experimenters and usually the participants in psychology research, there is the chance that 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

The Clever Hans case (Chapter 3, pp. 96–98) is often used to illustrate the influence of **experimenter bias** on the outcome of some study. Hans's trainer, knowing the outcome to the question "What is 3 times 3?," sent subtle head-nodding 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 rather emotionally involved in their research. It's not difficult to see how a desire to confirm some strongly held hypothesis might lead an unwary experimenter to behave 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. Robert 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 some judgment 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 (totally successful). All participants saw the same photos and made the same judgments. The independent variable was experimenter expectancy. Some experimenters were led to believe that most subjects would give people the benefit of the doubt and rate the pictures positively; other experimenters were told to expect negative ratings. Interestingly enough, 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 even being aware of it, are subtly shaping the responses 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, rats learn mazes faster for experimenters who *think* their animals have been bred for maze-running ability than for those expecting their rats to be "maze-dull" (Rosenthal & Fode, 1963b). The rats, of course, are randomly assigned to the experimenters and are equal in ability. The key factor here seems to be that experimenters expecting their rats to be "maze-bright" treat them better; for example, they handle them more, a behavior known to affect learning.

It should be noted that some of the Rosenthal research has been criticized on statistical grounds and for interpreting the results as being due to expectancy when they may have been due to something else. For example, Barber (1976) raised questions about the statistical conclusion validity of some of Rosenthal's work. In at least one study, according to Barber, 3 of 20 experimenters reversed the expectancy results, getting data the opposite of the expectancies created for them. Rosenthal omitted these experimenters from the analysis and obtained a significant difference for the remaining 17 experimenters. With all 20 experimenters included in the analysis, however, the difference disappeared. Barber also contends that, in the animal studies, some of the results occurred because experimenters simply fudged the data (e.g., misrecording maze errors). Another difficulty with the Rosenthal studies is that his procedures don't match what normally occurs in experiments; most experimenters test all of the participants in all conditions of the experiment, not just those participating in one of the conditions. Hence, Rosenthal's results might overestimate the amount of biasing that occurs.

Despite these reservations, 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 classification 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. With modern computer technology, participants can be shown photos on a screen and asked to make their responses with a key press while the experimenter is 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. Recall from Chapter 4 the study of rats in the radial maze, in which rat "macrochoices" and "microchoices" were confirmed by videotaping each animal's performance and defining those two constructs in terms of easily verifiable behaviors (Brown, 1992). Furthermore, computers make it easy to present instructions and stimuli to participants while also keeping track of data.

Experimenters can mechanize many procedures, to some degree at least, but the experimenter will be interacting with every participant nonetheless. Hence, it is important for experimenters to be given some training 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**.

Another strategy for controlling for experimenter bias is to use what is called a **double blind** procedure. This means simply that experimenters are kept in the dark (blind) about what to expect of participants in a particular testing session. As a result, neither the experimenters nor the participants know which condition is being tested—hence the designation "double." A double blind can be accomplished when the principal investigator sets up the experiment but a colleague (usually a graduate student) actually collects the data. Double blinds are not always possible, of course, as illustrated by the Dutton and Aron (1974) study you read about in Chapter 3. As you recall, female experimenters arranged to encounter men either on a suspension bridge swaying 230 feet over a river or on a solid bridge 10 feet over the same river. It would be a bit difficult to prevent those experimenters from knowing which condition of the study was being tested! On the other hand, many studies lend themselves to a procedure in which experimenters are blind to which condition is in effect. Research Example 6, which could increase the stock price of Starbucks, is a good example.

Research Example 6—Using a Double Blind

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" (I am writing this on an early Saturday morning, so I think I'll get it right). 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 into either a caffeine group or a decaf group (using Starbucks "house blends"). They were then given a standardized memory test on two different occasions, once at 8:00 a.m. and once at 4:00 p.m. The study was a double blind 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" (you will learn more about this concept in a few paragraphs). 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. In fact, most guessed incorrectly that they had been given regular coffee during one testing session and decaf at the other.

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 different groups, and this random assignment seemed to produce the desired equivalent groups—the groups were indistinguishable in terms of age, education level, and average daily intake of caffeine. Also, counterbalancing was used to insure that half of the seniors were tested first in the morning, then the afternoon, while the other half were tested in the sequence afternoon-morning.

The results? Time of day did not seem to affect a short-term memory task, but it had a significant effect on a more difficult longer-term task in which seniors learned some information, then had a 20-minute delay, then tried to recall the information, and then completed a recognition test for that same information. And 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 (max score = 16). Also, remember from Chapter 4 (pp. 137–138) 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, notice that I have included the standard deviations (*SD*).

Morning with caffeine →	11.8 ($SD = 2.9$)
Morning with decaf →	11.0 ($SD = 2.7$)
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 3:00 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 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" 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. 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 were 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 the 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 assembly of some 35 parts per relay. Western Electric produced about 7 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 were worsened. The example always mentioned concerned the infamous "12th test period" when workers were informed that the work week would increase from 42 to 48 hours per week, 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, 2003, p. 138)

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 some 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 things like 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). Her efforts contributed mightily to the high level of productivity.

A second problem with interpreting the relay data is a simple statistical problem. In the famous 12th period, productivity was recorded as output per week rather than output per hour, yet workers were putting in an extra 6 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 bonus money. Lastly, it could have been that 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).

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, such a picture also helps to maintain power at the level of management and impede efforts at unionization, 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, 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 very cooperative, to the point of persevering through repetitive and boring tasks, all in the name of psychological science. Furthermore, if participants can figure out the hypothesis, they may try to behave in such a way that confirms it. Orne 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, they no longer act naturally and it becomes 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?

Orne demonstrated how demand characteristics can influence a study's outcome by recruiting students for a so-called sensory deprivation experiment (Orne & Scheibe, 1964). He assumed that participants told that they were in such an experiment would expect the experience to be stressful and might respond accordingly.

This indeed occurred. Participants who sat for four hours in a small but comfortable room showed signs of stress *only* if (a) they signed a form releasing the experimenter from any liability in case anything happened to them, and (b) the room included a "panic button" that could be pressed if they felt too stressed by the deprivation. Control participants were given no release form to sign, no panic button to press, and no expectation that their senses were being deprived. They did not react adversely.

The possibility that demand characteristics are operating has an impact on decisions about whether to opt for between- or within-subject designs. Participants serving in all of the conditions of a study have a greater opportunity to figure out the hypothesis(es). Hence, demand characteristics are potentially more troublesome in within-subject designs than in between-subjects designs. For both types of designs, demand characteristics are especially devastating if they affect some conditions but not others, thereby introducing a confound.

Besides being good subjects (i.e., trying to confirm the hypothesis), participants 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 that they are in the condition of the study designed to reduce the chances that help will be offered. On the other hand, altruism is a valued, even heroic, behavior. The pressure to be a good subject and support the hypothesis pulls the participant toward nonhelping, 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).

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 (see Chapter 7, pp. 256–257). This procedure allows for a comparison between 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 just read (Research Example 6) used this kind of logic.

A second way to check for the presence of demand characteristics is to do what is sometimes called 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). This was accomplished in Research Example 6 by asking participants to guess whether they had been given caffeine in their coffee or not. Manipulation checks can also be done 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. Manipulation checks are also used to see if some procedure is producing the effect it is supposed to produce. For example, if some 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 that 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 and of privacy invasion in Chapter 3 (pp. 83–84).

Although I stated earlier that most research participants play the role of "good subjects," this is not uniformly true, and some 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 nonvolunteers 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 Group 1 was tested in the first half of the semester and Group 2 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 sign up first and the reluctant volunteers who wait as long as they can. Can you think of a way to control for this problem? If the concept "block randomization" 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 out this chapter, read Box 6.3, which concerns the ethical obligations of those participating 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 some responsibilities too.

Box 6.3

ETHICS—Research Participants Have Responsibilities Too

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

An article by Jim 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. They include

- ✓ Being responsible about scheduling by showing up for their appointments with researchers and arriving on time
- ✓ Being cooperative and acting professionally by giving their best and most honest effort
- ✓ Listening carefully to the experimenter during the informed consent and instructions phases and asking questions if they are not sure what to do
- ✓ Respecting any request by the researcher to avoid discussing the research with others until all the data have been collected
- ✓ Being 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 experimenters and participants. Korn's suggestion that participants take a more assertive role in making research more collaborative is a welcome one. This assertiveness, however, must be accompanied by enlightened experimenting that values and probes for the insights that 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.

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

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 be in another condition. With between-subjects designs, the main difficulty is creating groups that are essentially equivalent to each other on all factors except for the independent variable.

The Problem of 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 effectively 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.

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. Sequence effects include progressive effects (they gradually accumulate, as in fatigue) and carryover effects (one sequence of conditions might produce effects different from another sequence).

The Problem of Controlling Sequence Effects

Sequence effects are controlled by various counterbalancing procedures, all of which ensure that the different conditions are tested in more than one sequence. When participants serve in each condition of the study just once, complete (all possible sequences used) or partial (a sample of different sequences or a Latin square) counterbalancing will be used. When participants serve in each condition more than once, reverse counterbalancing or block randomization can be used. Asymmetric transfer can occur when carryover effects are present; such transfer reduces the effectiveness of counterbalancing.

Control Problems in Developmental Research

In developmental psychology, the major independent variable is age, a subject variable. If age is studied between subjects, the design is referred to as a cross-sectional design. 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 called a longitudinal design 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.

Problems with Biasing

The results of research in psychology can be biased by experimenter expectancy effects. These can lead the experimenter to treat participants in various conditions in different ways, making the results impossible to interpret. Such effects can be reduced by automating the procedures and using double blind control procedures. Participant bias also occurs. Participants might confirm the researcher's hypothesis if demand characteristics suggest to them the true purpose of a study or they might behave in unusual ways simply because they know they are in an experiment. Demand characteristics are usually controlled through varying degrees of deception and the extent of participant bias can be evaluated through the use of 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 the latter?
4. As a means of creating equivalent groups, when is matching most likely to be used?

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) block randomization were being used? How many sequences 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 how does it improve 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 and 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 are best cured by repeatedly exposing the person to the feared object and not allowing the person to escape until the person realizes that 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 spaced practice of verbal information will lead to greater retention than massed practice.
7. A clinician hypothesizes that people with an 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 8 hours, but start and end at different times).

Exercise 6.2—Constructing a Balanced Latin Square

A memory researcher wishes to compare long-term memory for a series of word lists as a function of whether the person initially studies either four lists or eight lists. Help the investigator in the planning stages of this project by constructing the two needed Latin squares, a 4 × 4 and an 8 × 8, using the procedure outlined in Table 6.2.

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, 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 the Self Tests:

✓ 6.1.

1. There is a minimum of two separate groups of subjects tested in the study, one group for each level of the IV; the problem of equivalent groups.
2. 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; sequence effects
2. With 6 levels of the IV, complete counterbalancing requires a minimum of 720 subjects (6 × 5 × 4 × 3 × 2 × 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 way that reflects bias.
3. If subjects know what is expected of them, they might be "good subjects" and not behave naturally.