CHAPTER SEVEN

Repeated Measures Designs

CHAPTER OUTLINE

Overview

Thus far we have considered experiments in which subjects participate in only one condition of the experiment. They are randomly assigned to one condition in the random groups and matched groups designs, or they are selected to be in one group in natural groups designs. These independent groups designs are powerful tools for studying the effects of a wide range of independent variables. There are times, however, when it is more effective to have each subject participate in all the conditions of an experiment. These designs are called **repeated measures designs** (or within-subjects designs). In an independent groups design, a separate group serves as a control for the group given the experimental treatment. In a repeated measures design, subjects *serve as their own controls* because they participate in both the experimental and control conditions.

Key Concept

We begin this chapter by exploring the reasons why researchers choose to use a repeated measures design. We then describe one of the central features of repeated measures designs. Specifically, in repeated measures designs, participants can undergo changes during the experiment as they are repeatedly tested. Participants may improve with practice, for example, because they learn more about the task or because they become more relaxed in the experimental situation. They also may get worse with practice—for example, because of fatigue or reduced motivation. These temporary changes are called *practice effects*.

We described in Chapter 6 that individual differences among participants cannot be eliminated in the random groups design, but they can be balanced by using random assignment. Similarly, the practice effects that participants experience due to repeated testing in the repeated measures designs cannot be eliminated. Like individual differences in the random groups design, however, practice effects can be balanced, or averaged, across the conditions of a repeated measures design experiment. When balanced across the conditions, practice effects are not confounded with the independent variable and the results of the experiment are interpretable.

Our primary focus in this chapter is to describe the techniques that researchers can use to balance practice effects. We also introduce data analysis procedures for repeated measures designs. We conclude the chapter with a consideration of problems that can arise in repeated measures designs.

Why Researchers Use Repeated Measures Designs

- Researchers choose to use a repeated measures design in order to (1) conduct an experiment when few participants are available, (2) conduct the experiment more efficiently, (3) increase the sensitivity of the experiment, and (4) study changes in participants' behavior over time.

Researchers gain several advantages when they choose to use a repeated measures design. First, repeated measures designs require fewer participants than an independent groups design, so these designs are ideal for situations in which only a small number of participants is available. Researchers who do experiments with children, the elderly, or special populations such as individuals with brain injuries frequently have a small number of participants available.

Researchers choose to use repeated measures designs even when sufficient numbers of participants are available for an independent groups design. The repeated measures designs often are more convenient and efficient. For example, Ludwig, Jeeves, Norman, and DeWitt (1993) conducted a series of experiments studying communication between the two hemispheres of the brain. The investigators measured how long it took participants to decide whether two briefly presented letters had the same name. The letters came from the set AaBb. Participants were to press the "match" key when the letters had the same name (AA, aa, Bb, bb) and the "no match" key when the letters had different names (AB, ab, Ab, aB). There were several different ways in which the pairs of letters were presented across four experiments, but there were two major conditions in these experiments. Either both letters were presented to one hemisphere (unilateral) or one letter of the pair was presented to each hemisphere (bilateral). Across four experiments, bilateral presentation led to faster response times than did unilateral presentation. In these experiments, two hemispheres were better than one!

Each trial in the Ludwig et al. (1993) experiment required only a few seconds to complete. The researchers could have tested separate groups of participants for the unilateral and bilateral conditions, but this approach would have been horribly inefficient. It would have taken more time to instruct participants regarding the nature of the task than it would have to do the task itself! A repeated measures design in which each participant was tested on both unilateral and bilateral trials provided the experimenters with a far more convenient and efficient way to answer their question about how the brain processes information.

Key Concept

Another important advantage of repeated measures designs is that they are generally more sensitive than an independent groups design. The **sensitivity** of an experiment refers to the ability to detect the effect of the independent variable even if the effect is a small one. Ideally, participants in a study respond similarly to an experimental manipulation. In practice, however, we know that people don't all respond the same way. This *error variation* can be due to variations in the procedure each time the experiment is conducted or to individual differences among the participants. An experiment is more sensitive when there is less variability in participants' responses within a condition of an experiment, that is, less error variation. In general, participants in a repeated measures design will vary within themselves less over the time of an experiment than participants in a random groups design will vary from other participants. Another way to say this is that there is usually more variation *between* people than there is *within* people. Thus, error variation will generally be less in a repeated measures design. The less error variation, the easier it is to detect the effect of an independent variable. The increased sensitivity of repeated measures designs is especially attractive to researchers who study independent variables that have small (hard-to-see) effects on behavior.

Researchers also choose to use a repeated measures design because some areas of psychological research require its use. When the research question involves studying changes in participants' behavior over time, such as in a learning experiment, a repeated measures design is needed. Further, whenever

BOX 7.1

REPEATED MEASUREMENTS AND THE REPEATED MEASURES DESIGN

It is important to distinguish among different situations in which researchers test participants repeatedly. For example, in Chapter 5 we saw that survey researchers administer surveys more than once to the same people in a longitudinal survey design in order to assess changes in respondents over time. In a repeated measures design *experiment*, researchers manipulate an independent variable to compare measures of participants' behavior in two or more conditions. The critical difference is that an independent variable is manipulated in the repeated measures design, but not in the longitudinal survey design.

Repeated testing also may be used when researchers investigate the reliability (consistency) of a measure. Researchers may obtain two (or more) measures of the same individuals in order to establish the reliability of a measure, called test–retest reliability (see Chapter 5). Repeated testing associated with the reliability of measurements differs from the repeated measures design. Only the repeated measures design involves an independent variable in which participants' responses are contrasted in different experimental conditions.

the experimental procedure requires that participants compare two or more stimuli relative to one another, a repeated measures design must be used. For example, a repeated measures design would have to be used if a researcher wanted to measure the minimum amount of light that must be added before participants could detect that a spot of light had become brighter. It would also be called for if a researcher wanted participants to rate the relative attractiveness of a series of photographs. Research areas such as psychophysics (illustrated by the light-detection experiment) and scaling (illustrated by the ratings of attractiveness) rely heavily on repeated measures designs. Journals such as *Perception & Psychophysics* and *Journal of Experimental Psychology: Human Perception and Performance* frequently publish results of experiments using repeated measures designs (see also Box 7.1).

The Role of Practice Effects in Repeated Measures Designs

- Repeated measures designs cannot be confounded by individual differences variables because the same individuals participate in each condition (level) of the independent variable.
- Participants' performance in repeated measures designs may change across conditions simply because of repeated testing (not because of the independent variable); these changes are called practice effects.
- Practice effects may threaten the internal validity of a repeated measures experiment when the different conditions of the independent variable are presented in the same order to all participants.
- There are two types of repeated measures designs (complete and incomplete) that differ in the specific ways in which they control for practice effects.

Defining Practice Effects

The repeated measures designs have another important advantage in addition to the ones we have already described. In a repeated measures design, the characteristics of the participants cannot confound the independent variable being manipulated in the experiment. The *same* participants are tested in all the conditions of a repeated measures design, so it is impossible to end up with brighter, healthier, or more motivated participants in one condition than in another condition. Stated more formally, *there can be no confounding by individual differences variables in repeated measures designs.* The absence of the potential for confounding by individual differences variables is a great advantage of the repeated measures designs. This does not mean, however, that there are no threats to the internal validity of experiments that are done using repeated measures designs.

One potential threat to internal validity arises because participants may change over time. The repeated testing of participants in the repeated measures design gives them practice with the experimental task. As a result of this practice, participants may get better and better at doing the task because they learn more about the task, or they may get worse at the task because of such factors as fatigue and boredom (see Figure 7.1). The changes participants undergo

FIGURE 7.1 There are both positive and negative effects of practicing a new skill. Repeating the same experience can lead to improvement, but it also can lead to fatigue, a decrease in motivation, and even boredom.

Key Concept

with repeated testing in the repeated measures designs are called **practice effects.** In general, practice effects should be balanced across the conditions in repeated measures designs so that practice effects "average out" across conditions. The key to conducting interpretable experiments using the repeated measures designs is learning to use appropriate techniques to balance practice effects. We will briefly introduce the two types of repeated measures designs before describing the use of specific balancing techniques.

Key Concept

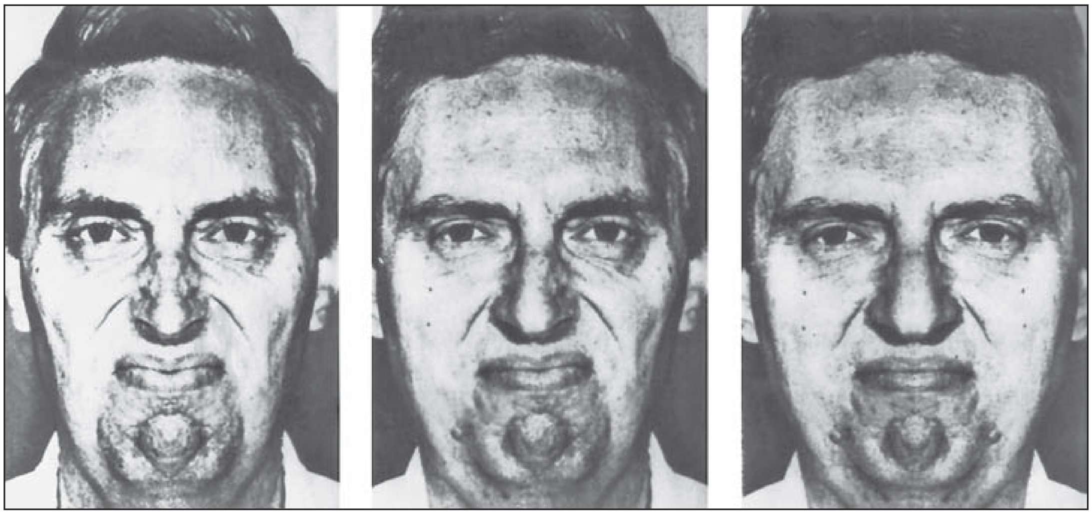
The two types of repeated measures designs are the complete and the incomplete design. The specific techniques for balancing practice effects differ for the two repeated measures designs, but the general term used to refer to these balancing techniques is **counterbalancing.** In the *complete design*, practice effects are balanced for *each* participant by administering the conditions to each participant several times, using different orders each time. Each participant can thus be considered a "complete" experiment. In the *incomplete design*, each condition is administered to each participant *only once*. The order of administering the conditions is varied across participants rather than for each participant, as is the case in the complete design. Practice effects in the incomplete design average out when the results are combined for all participants. This may seem a bit confusing at this point, but hopefully it will become clearer as we describe these types of designs more fully. Just keep in mind that a major goal when using a repeated measures design is to control for practice effects.

Balancing Practice Effects in the Complete Design

- Practice effects are balanced in complete designs within each participant using block randomization or ABBA counterbalancing.
- In block randomization, all of the conditions of the experiment (a block) are randomly ordered each time they are presented.
- In ABBA counterbalancing, a random sequence of all conditions is presented, followed by the opposite of the sequence.
- Block randomization is preferred over ABBA counterbalancing when practice effects are not linear, or when participants' performance can be affected by anticipation effects.

Research has shown that participants who view photographs depicting posed facial expressions of six basic human emotions (happiness, surprise, fear, sadness, anger, and disgust) can readily and accurately identify the expressed emotion. Sackeim, Gur, and Saucy (1978) used a repeated measures design to determine whether one side of our face expresses emotion more intensely than the other. They developed a photograph of a full face and a photograph of its mirror image. They then split both photographs down the middle making two composite photographs—one from the two versions of the left side of the face and one from the two versions of the right side. Illustrative photographs are presented in Figure 7.2. In the center is a photograph of a person expressing disgust. The two composite photographs made from the center photograph are presented on either side of the original. Does one of the two composites in Figure 7.2 look more disgusted than the other?

FIGURE 7.2 (a) Left-side composite, (b) original, and (c) right-side composite of the same face. The face is expressing disgust. (From Sackeim et al., 1978.)



Participants were shown slides of photographs like those in Figure 7.2 and were asked to rate each slide on a 7-point scale indicating the intensity of the expressed emotion. The slides were presented individually for 10 seconds, and participants were then given 35 seconds to make their rating. The critical independent variable in the experiment was the version of the photograph depicting one of the emotions (left composite, original, or right composite). Each participant rated 54 slides: 18 left composites, 18 originals, and 18 right composites.

Participants' ratings of emotional intensity were consistently higher for the left composite than for the right composite. Does this finding match your judgment that the face in panel (a) in Figure 7.2 appears more disgusted than the face in panel (c)? Sackeim et al. interpreted these findings in terms of hemispheric specialization of the brain. In general, the left hemisphere controls the right side of the body and the right hemisphere controls the left side of the body. Thus, the left composite reflects control by the right hemisphere, and the right composite reflects control by the left hemisphere. The higher ratings of emotional intensity for the left composite photographs suggest that the right hemisphere may be more heavily involved than the left hemisphere in the production of emotional expression.

The interpretation of the differences in the ratings depends critically on the order in which the slides were presented to participants. Consider what could happen if all the original versions were presented first, followed by all the right composites, then by all the left composites. If you imagine yourself in this experiment making a rating for each of the slides in this long sequence (over 40 minutes), you will get a sense of what we mean by practice effects. Surely your attention, motivation, and experience in rating the emotionality of photographs will change as you work through the sequence of slides. If you gave higher ratings for the slides shown at the end of this long sequence, your ratings may reflect the intensity of your own emotions of boredom and fatigue rather than the intensity of the emotions actually depicted in the photographs. To avoid this possibility, Sackeim et al. used balancing techniques specifically developed for use with the complete design in repeated measures experiments. By using these balancing techniques, they ensured that each of the three versions of the photographs was equally likely to appear at any point in the long series of slides.

In the complete design, participants are given each treatment enough times to balance practice effects for each participant. When the task is simple enough and not too time consuming (such as judging the emotional intensity of photographs), it is possible to give one participant several experiences with each treatment. In fact, in some complete designs, only one or two participants are tested, and each participant experiences literally hundreds of trials. More commonly, however, researchers use procedures like those used by Sackeim et al. That is, several participants are tested, and each participant is given each treatment only a relatively small number of times. Researchers have two choices in deciding how to arrange the order in which the treatments in a complete design are administered: block randomization and ABBA counterbalancing.

Block Randomization We introduced block randomization in Chapter 6 as an effective technique for assigning participants to conditions in the random groups design. *Block randomization* can also be used to order the conditions for each participant in a complete design. For instance, Sackeim et al. administered each of the three versions of their photographs (left composite, original, and right composite) 18 times to each participant. The sequence of trials shown in Table 7.1 illustrates how block randomization could be used to arrange the order of the three conditions in their experiment. The sequence of 54 trials is broken up into 18 blocks of 3 trials. Each block of trials contains the three conditions of the experiment in random order. In general, *the number of blocks in a block-randomized schedule is equal to the number of times each condition is administered, and the size of each block is equal to the number of conditions in the experiment.*

If a participant rated the photographs following the sequence in the block-randomized schedule shown in Table 7.1, it is unlikely that changes in the participant's attention, motivation, or experience with rating photographs would affect any one of the conditions more than any other. The practice effects can reasonably be expected to average out over the three experimental conditions. Determining the average position of each of the three conditions in the block-randomized sequence gives a rough indication of the balancing of practice effects. This can be done by summing the trial numbers on which each condition appears and dividing by 18. For instance, the original version of the photographs ("O") appeared on trials 1, 5, 8, 11, 13, 18, 21, 24, 27, 28, 33, 34, 39, 40, 44, 48, 49, and 53. The average position of the original photographs, therefore, was 27.6. The corresponding values for the left and right composite photographs are 27.7 and 27.2, respectively. That these average values are so similar tells us that any one version of the photographs was not more likely to appear at the beginning, middle, or end of the sequence of 54 trials.

Block randomization is effective in balancing practice effects, but each condition must be repeated several times before we can expect practice effects to

TABLE 7.1 BLOCK-RANDOMIZED SEQUENCE OF 54 TRIALS IN AN EXPERIMENT WITH THREE CONDITIONS ADMINISTERED 18 TIMES EACH

Trial	Conditions		Trial	Conditions
1	O	First Block	28	O
2	L		29	L
3	R		30	R
4	R		31	R
5	O		32	L
6	L		33	O
7	R		34	O
8	O		35	R
9	L		36	L
10	L		37	L
11	O		38	R
12	R		39	O
13	O		40	O
14	L		41	R
15	R		42	L
16	R		43	R
17	L		44	O
18	O		45	L
19	R		46	R
20	L		47	L
21	O		48	O
22	L		49	O
23	R		50	R
24	O		51	L
25	R		52	R
26	L		53	O
27	O		54	L

Note: The conditions are the three versions of the photographs used by Sackeim et al. (1978): L = left composite, O = original, R = right composite.

average out. We should not expect practice effects to be balanced after two or three blocks—any more than we would expect sample sizes of two or three in the random groups design to result in comparable groups. Fortunately, a technique is available to balance practice effects when it is not possible to administer each condition often enough for the averaging process of block randomization to work effectively.

ABBA Counterbalancing In its simplest form, ABBA counterbalancing can be used to balance practice effects in the complete design with as few as two administrations of each condition. *ABBA counterbalancing* involves presenting the conditions in one sequence (i.e., A then B) followed by the opposite of that

TABLE 7.2 ABBA COUNTERBALANCED SEQUENCE OF TRIALS IN AN EXPERIMENT WITH THREE CONDITIONS (LEFT COMPOSITE, ORIGINAL, AND RIGHT COMPOSITE)

	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6
Condition:	Left	Original	Right	Right	Original	Left
Practice effect (linear)	+0	+1	+2	+3	+4	+5
Practice effect (nonlinear)	0	+6	+6	+6	+6	+6

same sequence (i.e., B then A). Its name describes the sequences when there are only two conditions (A and B) in the experiment, but ABBA counterbalancing is not limited to experiments with just two conditions. Sackeim et al. could have presented the versions of their photographs according to the ABBA sequence outlined in the top row of Table 7.2 labeled "Condition." Note that in this case it literally would be ABCCBA since there are three conditions. The order of the three conditions on the first three trials is simply reversed for trials 4 to 6.

ABBA counterbalancing is appropriately used only when practice effects are linear. If practice effects are linear, the same amount of practice effects is added to or subtracted from performance on each successive trial. The row of Table 7.2 labeled "Practice effect (linear)" illustrates how ABBA counterbalancing can balance practice effects. In this example, one "unit" of hypothetical practice effects is added to performance on each trial. Because there would be no practice effect associated with the first trial, the amount of practice added to Trial 1 in the table is zero. Trial 2 has one unit of hypothetical effects added because of participants' experience with the first trial; in Trial 3 there are two units added because of participants' experience with two trials, and so on.

We can get an idea of the influence of practice effects by adding the values for each condition. For example, the left composite condition gets the least (0) and the greatest (+5) influence from practice effects; the right composite condition gets two intermediate amounts (+2 and +3). The sum of the hypothetical practice effects is +5 for both conditions. (What would the sum of the practice effects be for the original condition?) The ABBA cycle can be applied with any number of conditions, but there must be an even number of repetitions of each condition. ABBA counterbalancing balances practice effects even more effectively with larger numbers of repetitions of the cycle. Usually, however, ABBA counterbalancing is used when the number of conditions and the number of repetitions of each condition are relatively small.

Although ABBA counterbalancing provides a simple and elegant means to balance practice effects, it is not without limitations. For example, ABBA counterbalancing is ineffective when practice effects for a task are not linear. This is illustrated in the last row of Table 7.2, labeled "Practice effect (nonlinear)." Nonlinear practice effects can occur when participants' performance changes dramatically after exposure to one or more trials. In this example, the left composite receives a total of only six hypothetical units of practice effects, and the other two conditions receive a total of 12 units each. When practice effects involve abrupt initial changes followed by little change thereafter, researchers often ignore performance on the early trials and wait until the practice effects reach a "steady state." Reaching a steady state is likely to take several repetitions of each condition, so researchers tend to use block randomization to balance practice effects in these situations.

ABBA counterbalancing is also ineffective when anticipation effects can occur. *Anticipation effects* occur when a participant develops expectations about which condition should occur next in the sequence. The participant's response to that condition may then be influenced more by this expectation than by the actual experience of the condition itself. For example, consider a time-perception experiment in which the participant's task is to estimate the length of time that has passed between a signal presented on a computer screen indicating the start of an interval and another signal indicating the end of the interval. (Of course, participants have to be prevented somehow from marking off time during the interval by counting or rhythmically tapping.) If the time intervals in such an experiment are 12, 24, and 36 seconds, then one possible ABBA sequence of conditions could be 12-24-36-36-24-12. If this cycle were repeated several times, participants probably would recognize the pattern and expect a series of increasing and then decreasing intervals. Their time estimates might soon begin to reflect this pattern rather than their perception of each independent interval. If anticipation effects are likely, block randomization should be used rather than ABBA counterbalancing.

Balancing Practice Effects in the Incomplete Design

- Practice effects are balanced *across* subjects in the incomplete design rather than for each subject, as in the complete design.
- The rule for balancing practice effects in the incomplete design is that each condition of the experiment must be presented in each ordinal position (first, second, etc.) equally often.
- The best method for balancing practice effects in the incomplete design with four or fewer conditions is to use all possible orders of the conditions.
- Two methods for selecting specific orders to use in an incomplete design are the Latin Square and random starting order with rotation.
- Whether using all possible orders or selected orders, participants should be randomly assigned to the different sequences.

In the incomplete design, each participant is given each treatment *only once*. The results for any one participant, therefore, cannot be interpreted because the levels of the independent variable for each participant are perfectly confounded with the order in which those levels were presented. For instance, the first participant in an incomplete design experiment might be tested first in the experimental condition (E) and second in the control condition (C). Any differences in the participant's performance between the experimental and control conditions could be due to the effect of the independent variable *or* to the practice effects resulting from the EC order. To break this confounding of the order of conditions and the independent variable, we can administer different orders of the conditions to different participants. For example, we could administer the conditions of our incomplete design experiment to a second participant in the CE order, testing the control condition first and the experimental condition second. In this way, we could balance the effects of order across the two conditions using two participants instead of one.

To illustrate the techniques for balancing practice effects in the incomplete design, we will use a repeated-measures experiment from the field of health psychology that investigated the effects of aerobic exercise on participants' moods (Hansen, Stevens, & Coast, 2001). The purpose of the study was to determine the time interval(s) of exercise required for mood improvements, and the researchers compared 30 minutes of quiet resting (0 exercise) to 10, 20, and 30 minutes of exercise. The exercise consisted of riding a stationary ergometric bicycle that allowed heart rate (HR) monitoring. During the exercise sessions, a warmup period was used to reach a target HR of moderate exercise intensity, then participants cycled for the required amount of time in the trial while maintaining that heart rate. Before exercise and after a cooldown period (following exercise), participants completed a mood inventory to assess their mood "at that moment." Each female participant was tested in each of the four conditions, with testing sessions on the same day of the week, one week apart for four consecutive weeks. Participants were randomly assigned to an order of the four conditions.

Before describing the technique that can be used to balance practice effects for an independent variable in the incomplete design, we will take a brief look at the results of the Hansen et al. study. The dependent variable in this study was the difference between participants' mood ratings before exercise and after exercise (and before and after resting in the 0 exercise condition). The researchers examined changes in depression, anxiety, anger, fatigue, and confusion (e.g., feeling overwhelmed), and a positive mood state of vigor. Overall, results indicated that exercise improved vigor and decreased confusion, fatigue, and total negative mood (a sum of mood scores). How much exercise was needed to see these effects? Analyses indicated that these improvements occurred with just 10 minutes of exercise! With 20 minutes of exercise participants experienced further improvements in feelings of confusion; no additional mood gains were seen when participants reached 30 minutes of exercise. Hansen et al. (2001) concluded that, in conjunction with recommendations regarding fitness (e.g., Centers for Disease Control), "to experience positive fitness and health benefits, healthy adults should participate in a total of thirty minutes of moderate physical exercise daily, accumulated in short bouts throughout the day" (p. 267).

We turn our attention now to the balancing techniques that are used in the incomplete design. In an incomplete design it is essential that practice effects be balanced by varying the order in which the conditions are presented. The general rule for balancing practice effects in the incomplete design is a simple one: *Each condition of the experiment must appear in each ordinal position (1st, 2nd, 3rd, etc.) equally often*. Several techniques are available for satisfying this general rule. These techniques differ in what additional balancing they accomplish, but so long as the techniques are properly used, the basic rule will be met and the experiment will be interpretable. That is, if appropriate balancing is carried out, then we will be in a position to determine whether the independent variable, not practice effects, influenced the participants' behavior.

All Possible Orders The preferred technique for balancing practice effects in the incomplete design is to use all possible orders of the conditions. Each participant is randomly assigned to one of the orders. With only two conditions there are only two possible orders (AB and BA); with three conditions there are six possible orders (ABC, ACB, BAC, BCA, CAB, CBA). In general, there are $N!$ (which is read "N factorial") possible orders with N conditions, where $N!$ equals $N(N - 1)\ (N - 2) \ldots (N - [N - 1])$. As we just saw, there are six possible orders with three conditions, which is 3! ($3 \times 2 \times 1 = 6$). The number of required orders increases dramatically with increasing numbers of conditions. For instance, for five conditions there are 120 possible orders, and for six conditions there are 720 possible orders. Because of this, the use of all possible orders is usually limited to experiments involving four or fewer conditions.

Because there were four conditions in the Hansen et al. (2001) exercise experiment, 24 sequences would be required to obtain all possible orders of conditions. These sequences (orders of conditions) are presented in the left half of Table 7.3. Using all possible orders certainly meets the general rule of ensuring that all conditions appear in each ordinal position equally often. The first ordinal position shows this balancing most clearly: The first six sequences begin with the 0 exercise condition, and each of the next six sets of sequences begins with one of the three exercise conditions. The same pattern applies at each of the four ordinal positions. For example, the "0" condition also appears six times in the second ordinal position, six times in the third ordinal position, and six times in the fourth ordinal position. The same is true for the 10-, 20-, and 30-minute exercise conditions.

TABLE 7.3 ALTERNATIVE TECHNIQUES TO BALANCE PRACTICE EFFECTS IN AN INCOMPLETE REPEATED MEASURES DESIGN EXPERIMENT WITH FOUR CONDITIONS

All Possible Orders								Selected Orders: Latin Square				Selected Orders: Random Starting Order with Rotation			
Ordinal Position				Ordinal Position				Ordinal Position				Ordinal Position			
1st	2nd	3rd	4th	1st	2nd	3rd	4th	1st	2nd	3rd	4th	1st	2nd	3rd	4th
0	10	20	30	20	0	10	30	0	10	20	30	10	20	30	0
0	10	30	20	20	0	30	10	10	30	0	20	20	30	0	10
0	20	10	30	20	10	0	30	30	20	10	0	30	0	10	20
0	20	30	10	20	10	30	0	20	0	30	10	0	10	20	30
0	30	10	20	20	30	0	10								
0	30	20	10	20	30	10	0								
10	0	20	30	30	0	10	20								
10	0	30	20	30	0	20	10								
10	20	0	30	30	10	0	20								
10	20	30	0	30	10	20	0								
10	30	0	20	30	20	0	10								
10	30	20	0	30	20	10	0								

Note: The four conditions are identified using the time of exercise in the Hansen et al. (2001) experiment: 0 exercise, 10 minutes, 20 minutes, and 30 minutes.

There is one other issue that must be addressed in deciding to use all possible orders. For this technique to be effective, it is essential that at least one participant be tested with each of the possible orders of the conditions. That is, at least one participant should receive the 0-10-20-30 order, at least one should receive the 0-10-30-20 order, and so on. Therefore, the use of all possible orders requires at least as many participants as there are possible orders. Thus, if there are four conditions in the experiment, at least 24 participants are needed (or 48, or 72, or some other multiple of 24). This restriction makes it very important that a researcher has a good idea of the number of potential participants available before testing the first participant.[1]

Selected Orders We have just described the preferred method for balancing practice effects in the incomplete design, all possible orders. There are times, however, when the use of all possible orders is not practical. For example, if we wanted to use the incomplete design to study an independent variable with seven levels, we would need to test 5,040 participants if we used all possible orders—one participant for each of the possible orders of the seven conditions (7! orders). We obviously need some alternative to using all possible orders if we are to use the incomplete design for experiments with five or more conditions.

Practice effects can be balanced by using just some of all the possible orders. The number of selected orders will always be equal to some multiple of the number of conditions in the experiment. For example, to do an experiment with one independent variable with seven levels, we need to select 7, 14, 21, 28, or some other multiple of seven orders to balance practice effects. The two basic variations of using selected orders are illustrated in Table 7.3. To allow you to compare the types of balancing more directly, we have illustrated the techniques for selected orders with the four-level independent variable from the Hansen et al. (2001) experiment that we described in the previous section.

The first type of balancing using selected orders is called the Latin Square. In a *Latin Square*, the general rule for balancing practice effects is met. That is, each condition appears at each ordinal position once. For example, just to the right of the center of Table 7.3, we can see that in the Latin Square, condition "0" appears exactly once in the first, second, third, and fourth ordinal positions. This is true for each condition. Additionally, in a Latin Square each condition precedes and follows each other condition exactly once. Examination of the Latin Square in Table 7.3 shows that the order "0–10" appears once, as does the order "10–0." The order "10–20" appears once, as does the order "20–10," and so on, for every combination of conditions. (The procedure for constructing a Latin Square is described in Box 7.2.)

The second balancing technique using selected orders requires you to begin with a random order of the conditions and to rotate this sequence systematically

[1]The number of participants ($N = 14$) in the Hansen et al. (2001) exercise study made it impossible for them to use all possible orders. Instead, they identified a random order of conditions for each participant. This leaves open the possibility that practice effects were not completely balanced in their design. For example, if the 10-minute exercise period was more often last in the sequence, participants' mood improvement may have been due to relief over a shorter exercise period, not the exercise itself.

BOX 7.2

HOW TO CONSTRUCT A LATIN SQUARE

A simple procedure for constructing a square *with an even number (N) of conditions* is as follows:

1 Randomly order the conditions of the experiment.
2 Number the conditions in your random order 1 through *N*.

Thus, if you had $N = 4$ conditions (A, B, C, D) and the random order (from Step 1) was B, A, D, C, then B = 1, A = 2, D = 3, C = 4.

3 To generate the first row (first order of conditions), use the rule

1, 2, N, 3, $N - 1$, 4, $N - 2$, 5, $N - 3$, 6, etc.

In our example, this would yield 1, 2, 4, 3.

4 To generate the second row (second order of conditions), add 1 to each number in the first row but with the understanding that 1 added to $N = 1$.

We would then have 2, 3, 1, 4.

5 The third row (third order of conditions) is generated by adding 1 to each number in the second row and again $N + 1 = 1$.

The third row would be 3, 4, 2, 1.

6 A similar procedure is carried out for each successive row.

Can you construct the fourth row in this 4×4 square?

7 Assign the conditions to their corresponding numbers as determined in Step 2.

The Latin Square for this example would be

B A C D
A D B C
D C A B
C B D A

If there is an odd number of conditions, then two squares must be constructed. The first can be made according to the rule given above for even-numbered squares. The second square is generated by reversing the rows in the first square. For example, assume $N = 5$ and the first row of the first square is B A E C D. The first row of the second square would then be D C E A B. The two squares are joined to make an $N \times 2N$ square. In either case, even or odd, subjects should be assigned randomly to the rows of the square. Thus, you must have available at least as many subjects as there are multiples of rows. (Procedures for selecting or constructing Latin Squares are also described in Winer, Brown, and Michels [1991, pp. 674–679].)

with each condition moving one position to the left each time (see the example on the right in Table 7.3). Using a random starting order with rotation effectively balances practice effects because, like the Latin Square, each condition appears in each ordinal position. However, the systematic rotation of the sequences means that each condition always follows and always precedes the *same* other conditions (e.g., 30 always comes after 20 and before 0), which is not like the Latin Square technique. The simplicity of the random starting order with rotation technique and its applicability to experiments with more than four conditions are its primary advantages.

The use of all possible orders, Latin Squares, and random starting orders with rotation are equally effective in balancing practice effects because all three techniques ensure that each condition appears in each ordinal position equally often. Regardless of which technique one uses to balance practice effects, the sequences of conditions should be fully prepared prior to testing the first participant, and participants should be randomly assigned to these sequences.

Data Analysis of Repeated Measures Designs

Describing the Results

- Data analysis for a complete design begins with computing a summary score (e.g., mean, median) for each participant.
- Descriptive statistics are used to summarize performance across all participants for each condition of the independent variable.

After checking the data for errors and outliers, the first step in analyzing a repeated measures experiment is to summarize participants' performance in each condition of the experiment. In random groups designs, this means simply listing the scores of the participants tested in each of the conditions of the experiment and then summarizing these scores with descriptive statistics such as the mean and standard deviation. In an incomplete repeated measures design, each participant provides one score in each condition, but it is still relatively straightforward to summarize the scores for each condition. In doing so, you need to be careful as you "unwind" the various orders in which the participants were tested to be sure participants' scores are listed with the correct condition. Once all the scores for each condition have been listed together, means and standard deviations can be computed to describe performance in each condition.

An additional step needs to be taken when analyzing a complete repeated measures design. You first must compute a score for each participant in each condition before you begin to summarize and describe the results. This additional step is necessary because each participant is tested in each condition more than once in a complete design. For example, five participants were tested in a time-perception experiment done as a classroom demonstration of a complete repeated measures design. The purpose of the experiment was not to test the accuracy of participants' time estimates compared with the actual interval lengths. Instead, the purpose of the experiment was to determine whether participants' estimates of time increased systematically with increasing lengths. In other words, could participants discriminate between intervals of different lengths?

Each participant in the experiment was tested six times on each of four interval lengths (12, 24, 36, and 48 seconds). Block randomization was used to determine the order in which the intervals were presented. Thus, each participant provided 24 time estimates, six estimates for each of the four interval lengths. Any one of the six estimates for a given time interval is contaminated by practice effects, so some measure that combines information across the six estimates is needed. Typically, the mean across the six estimates for each interval would be calculated for each participant to provide a single estimate of performance in each condition. As you may remember, however, extreme scores can influence the mean; it is quite possible that participants gave extreme estimates of the time intervals for at least one of the six tests of each interval. Thus, for this particular set of data, the median of the six estimates probably provides the best measure to reflect the participants' estimates of the time intervals. These median estimates (rounded to the nearest whole number) are listed in Table 7.4. (You may be used to seeing the mean and median as descriptive statistics summarizing a

TABLE 7.4 DATA MATRIX TABLE FOR A REPEATED MEASURES DESIGN EXPERIMENT

Data Matrix				
	Interval Length			
Participant	12	24	36	48
1	13	21	30	38
2	10	15	38	35
3	12	23	31	32
4	12	15	22	32
5	16	36	69	60
Mean (*SD*)	12.6 (2.0)	22.0 (7.7)	38.0 (16.3)	39.4 (10.5)

Note: Each value in the table represents the median of the participants' six responses at each level of the interval length variable. The means in the bottom row are the averages of the medians (from six responses made by the five participants at each interval length).

group's performance; however, as this example illustrates, these summary statistics also can be used to represent one *person's* performance when that performance is an "average" across trials or tests.)

Once an individual score for each participant in each condition has been obtained, the next step is to summarize the results across participants, using appropriate descriptive statistics. The mean estimate and standard deviation (*SD*) for each of the four intervals are listed in the row labeled "Mean (*SD*)" in Table 7.4. Even though the data for only five participants are included in the table, these mean estimates indicate that participants appear to have discriminated between intervals of different lengths, at least for intervals up to 36 seconds.

As we mentioned in Chapter 6, it is a good idea to include measures of effect size when describing the results of an experiment. A typical measure of effect size for a repeated measures design is the strength of association measure called eta squared (η^2). The value of eta squared for the time-perception experiment was .80. This value indicates that a large proportion of variation in participants' time estimates can be accounted for by the independent variable of interval length. You can find more information about the calculation of effect sizes and their interpretation in Chapters 11 and 12. In Chapter 12 we illustrate how to calculate eta squared using the data found in Table 7.4.

Confirming What the Results Reveal

- The general procedures and logic for null hypothesis testing and for confidence intervals for repeated measures designs are similar to those used for random groups designs.

Data analysis for experiments using repeated measures designs involves the same general procedures we described in Chapter 6 for the analysis of random groups design experiments. Researchers use null hypothesis testing and confidence intervals to make claims about whether the independent variable produced an effect on behavior. We will use the time-perception experiment to illustrate how researchers confirm what the data reveal when they use repeated measures designs.

The focus of the analysis of the time-perception experiment was on whether the participants could discriminate intervals of different lengths. We cannot make the claim that participants were able to discriminate intervals of varying lengths until we know that the mean differences in Table 7.4 are greater than would be expected on the basis of error variation alone. That is, even though it may *appear* that participants were able to discriminate between the different intervals, we do not know if their performance was different from that which would occur by chance. Thus, we must consider using analytical tools of null hypothesis testing and the construction of confidence intervals to help us make a decision about the effectiveness of the independent variable.

One distinctive characteristic of the analysis of repeated measures designs is the way in which error variation is estimated. We described in Chapter 6 that for the random groups design, individual differences among participants within the groups provides an estimate of error variation. In repeated measures designs, however, differences among participants are not just balanced—they are actually eliminated from the analysis. The ability to eliminate systematic variation due to participants in repeated measures designs makes these designs generally more sensitive than random groups designs. The source of error variation in the repeated measures designs is the differences in the ways the conditions affect different participants.

STRETCHING EXERCISE

For this exercise you are to compute the mean for each level of the independent variable in this complete repeated measures design. You must first compute a summary score for each participant in each condition before you summarize and describe the results for the three conditions.

In a perception experiment, three participants were tested for their ability to identify complex visual patterns. On each presentation the participants briefly viewed a complex pattern (target), followed by a test with a set of four patterns (the target and three other similar patterns). Their task was to pick out the target pattern from the set. The independent variable was the delay between the target and the test, with three levels: 10s, 30s, 50s. On each of six trials, participants made 50 judgments at one level of the independent variable. The table shows the ABBA counterbalanced sequence of trials for each participant on each trial. The values in parentheses represent the number of errors (the dependent variable) made by each participant on each trial (50 max.). Use this table to describe the effect of the delay independent variable on the number of errors.

Participant	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6
1	30s (9)	50s (6)	10s (2)	10s (6)	50s (10)	30s (3)
2	50s (10)	30s (6)	10s (2)	10s (4)	30s (8)	50s (8)
3	10s (1)	50s (6)	30s (7)	30s (3)	50s (8)	10s (3)

The fact that error variation is estimated differently in a repeated measures design than it is in an independent groups design means that the calculation of the *t*-test and *F*-test used in null hypothesis testing also differs. Similarly, there is change in the way that confidence intervals are calculated. In Chapter 12 we use the data in Table 7.4 to show how both the *F*-test and confidence intervals are used in decision making as part of a repeated measures design. The null hypothesis for an analysis of the data in Table 7.4 is that the population means, estimated by the sample means, are the same across interval-length conditions. Having carried out an analysis of variance for these data (see Chapter 12), we can tell you that the probability associated with the *F*-test for the effect of interval length was $p = .0004$. Because this obtained probability is less than the conventional level of significance (.05), the effect of the interval length variable was statistically significant. Based on this outcome, we can make the claim that participants' time estimates did differ systematically as a function of interval length. We already know from our calculation of the effect size (eta squared = .80) that it represents a large effect.

In Chapter 12 we used the same data to calculate .95 confidence intervals for the means seen in Table 7.4. The confidence intervals (in seconds) for the four conditions are (12) 5.4–19.8; (24) 14.8–29.2; (36) 30.8–45.2; (48) 32.2–46.6. As you learned in Chapter 6 (see also Box 11.5), when intervals do not overlap, we can claim that the population means estimated by the sample means are different. Does an inspection of these intervals tell you which means would be judged to be different? A convenient way to examine the relationship among confidence intervals is to plot them in a graph. For example, take a look at Figure 12.2 in Chapter 12, in which the intervals presented here are plotted around the sample means obtained in the time estimation experiment.

The Problem of Differential Transfer

- Differential transfer occurs when the effects of one condition persist and influence performance in subsequent conditions.
- Variables that may lead to differential transfer should be tested using a random groups design because differential transfer threatens the internal validity of repeated measures designs.
- Differential transfer can be identified by comparing the results for the same independent variable when tested in a repeated measures design and in a random groups design.

Researchers can overcome the potential problem of practice effects in repeated measures designs by using appropriate techniques to balance practice effects. There is a much more serious potential problem that can arise in repeated measures designs that is known as differential transfer (Poulton, 1973, 1975, 1982; Poulton & Freeman, 1966). **Differential transfer** arises when performance in one condition differs depending on the condition that precedes it.

Key Concept

Consider a problem-solving experiment in which two types of instructions are being compared in a repeated measures design. One set of instructions (A) is expected to enhance problem solving, whereas the other set of instructions (B) serves as the neutral control condition. It is reasonable to expect that participants tested in the order AB will be unable or unwilling to abandon the approach outlined in the A instructions when they are supposed to be following the B instructions. Giving up the "good thing" participants had under instruction A would be the counterpart of successfully following the admonition "Don't think of pink elephants!" When participants fail to give up the instruction from the first condition (A) while they are supposed to be following instruction B, any difference between the two conditions is reduced. For those participants, after all, condition B was not really tried. The experiment becomes a situation in which participants are tested in an "AA" condition, not an "AB" condition.

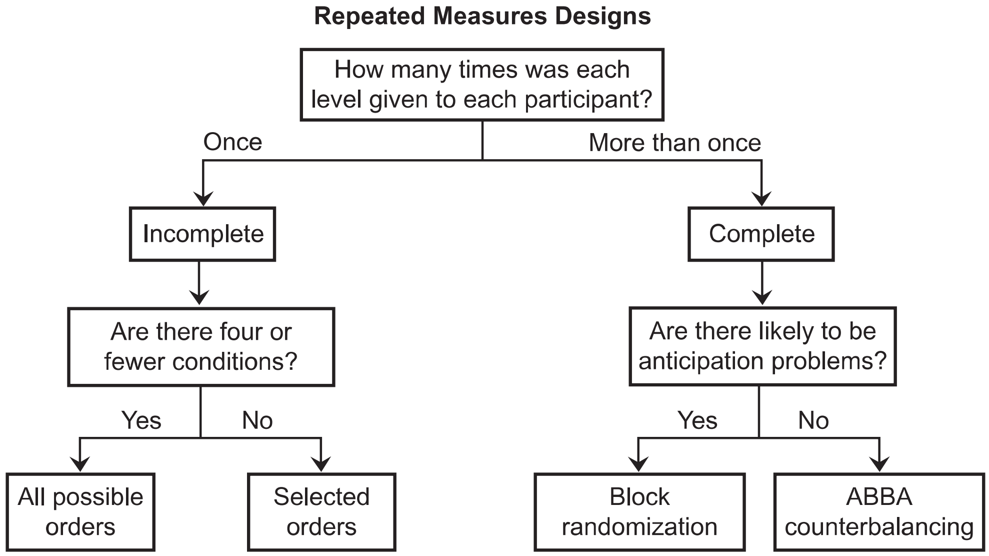
In general, the presence of differential transfer threatens internal validity because it becomes impossible to determine if there are true differences between the conditions. It also tends to underestimate differences between the conditions and thereby reduces the external validity of the findings. *Therefore, when differential transfer could occur, researchers should choose an independent groups design.* Differential transfer is sufficiently common with instructional variables to advise against the use of repeated measures designs for these studies (Underwood & Shaughnessy, 1975). Unfortunately, differential transfer can arise in any repeated measures design. For instance, the effect of 50 units of marijuana may be different if administered after the participant has received 200 units than if administered after the participant has received the placebo (e.g., if the participant has an increased tolerance for marijuana after receiving the 200 dose). There are ways, however, to determine whether differential transfer is likely to have occurred.

The best way to determine whether differential transfer is a problem is to do two separate experiments (Poulton, 1982). The same independent variable would be studied in both experiments, but a random groups design would be used in one experiment and a repeated measures design in the other. The random groups design cannot possibly involve differential transfer because each participant is tested in only one condition. If the experiment using a repeated measures design shows the same effect of the independent variable as that shown in the random groups design, then there has likely been no differential transfer. If the two designs show different effects for the same independent variable, however, differential transfer is likely to be responsible for producing the different outcome in the repeated measures design. When differential transfer does occur, the results of the random groups design should be used to provide the best description of the effect of the independent variable.

SUMMARY

Repeated measures designs provide an effective and efficient way to conduct an experiment by administering all the conditions in the experiment to each participant (see Figure 7.3). Repeated measures designs are useful when only very few participants are available or when an independent variable can be

FIGURE 7.3 In this chapter we introduced repeated measures designs and methods for counterbalancing.



studied most efficiently by testing fewer participants several times. Repeated measures designs are generally more sensitive experiments. Finally, particular areas of psychological research (e.g., psychophysics) may require the use of repeated measures designs.

For any repeated measures design experiment to be interpretable, however, practice effects must be balanced. Practice effects are changes that participants undergo because of repeated testing. In a complete repeated measures design, practice effects are balanced for each participant. Block randomization and ABBA counterbalancing can be used to balance practice effects in a complete repeated measures design. ABBA counterbalancing should not be used, however, if practice effects are expected to be nonlinear or if anticipation effects are likely.

In an incomplete repeated measures design, each participant receives each treatment only once, and the balancing of practice effects is accomplished across participants. Techniques for balancing practice effects in an incomplete repeated measures design involve either the use of all possible orders or selected orders (the Latin Square and rotation of a random starting order).

The process of data analysis of the results of repeated measures designs is essentially the same as that for analyzing the results of random groups designs. An added step for the complete repeated measures design is that each participant's scores first must be summarized within each condition. The data are examined for errors and then summarized using descriptive statistics such as the mean, standard deviation, and measures of effect size. Null hypothesis testing and confidence intervals are used to make claims that the independent variable has produced an effect on behavior.

The most serious problem in any repeated measures design is differential transfer—when performance in one condition differs depending on which condition it follows. Procedures for detecting the presence of differential transfer are available, but there is little that can be done to salvage a study in which it occurs.

Key Concepts

repeated measures designs 226
sensitivity 227
practice effects 230
counterbalancing 230
differential transfer 243

Review Questions

1 Describe what is balanced in a random groups design and what is balanced in a repeated measures design.
2 Briefly describe four reasons why researchers would choose to use a repeated measures design.
3 Define sensitivity and explain why repeated measures designs are often more sensitive than random groups designs.
4 Distinguish between a complete design and an incomplete design for repeated measures designs.
5 What options do researchers have in balancing practice effects in a repeated measures experiment using a complete design?
6 Under what two circumstances would you recommend against the use of ABBA counterbalancing to balance practice effects in a repeated measures experiment using a complete design?
7 State the general rule for balancing practice effects in repeated measures experiments using an incomplete design.
8 Briefly describe the techniques that researchers can use to balance practice effects in the repeated measures experiments using an incomplete design. Identify which of these techniques is preferred and explain why.
9 Explain why an additional initial step is required to summarize the data for an experiment involving a complete repeated measures design.
10 Describe how researchers can determine if differential transfer has occurred in a repeated measures experiment.

CHALLENGE QUESTIONS

1 The following problems represent different situations in the repeated measures designs in which practice effects need to be balanced.
   A Consider a repeated measures experiment using a complete design involving one independent variable. The independent variable in the experiment is task difficulty with three levels (Low, Medium, and High). You are to prepare an order for administering the conditions of this experiment so that the independent variable is balanced for practice effects. You are first to use block randomization to balance practice effects and then to use ABBA counterbalancing to balance practice effects. Each condition should appear twice in the order you prepare. (You can use the first row of the random number table (Table A.1) in the Appendix to determine your two random orders for block randomization.)
   B Consider a repeated measures experiment using an incomplete design. The independent variable in the experiment is the font size in which a paragraph has been printed, and there are six levels (7, 8, 9, 10, 11, and 12). Present a table showing how you would determine the order of administering the conditions to the first six participants of the experiment. Be sure that practice effects are balanced for these participants.
2 The pursuit rotor is a test of perceptual-motor coordination. It involves a turntable with a disk about the size of a dime embedded in it. The participant is given a pointer and is asked to keep the pointer on the disk while the turntable is

rotating. The dependent variable is the percentage of time on each trial that the participant keeps the pointer on the disk. Learning on this task is linearly related to trials over many periods of practice, and the task generally takes a long time to master. A researcher wants to study the influence of time of day on the performance on this task with four different times (10 A.M., 2 P.M., 6 P.M., and 10 P.M.). The participants will receive a constant number of trials under each of the four conditions, and participants will be tested in one condition per day over four consecutive days.

- **A** What design is being used for the time-of-day variable in this experiment?
- **B** Prepare a Latin Square to balance practice effects across the conditions of this experiment.
- **C** The researcher decides to use all possible orders to balance practice effects. The researcher assigns each participant to one of the 24 possible orders of the conditions. Which experimental design is included when you look only at the first condition to which each participant was assigned?
- **D** How could the researcher test whether differential transfer occurred when all possible orders were used to balance practice effects?

3 The following table represents the order of administering the conditions to participants in a repeated measures experiment using an incomplete design in which the independent variable was the loudness of a tone to be detected by the participants while they were concentrating on another task. The three tones were extremely soft (ES), very soft (VS), and soft (S). The values in parentheses represent the number of times each participant detected the tone in each condition. Use this table, when necessary, to answer questions that follow.

Participant	Order of Conditions		
1	ES (2)	VS (9)	S (9)
2	VS (3)	S (5)	ES (7)
3	S (4)	ES (3)	VS (5)
4	ES (6)	S (10)	VS (8)
5	VS (7)	ES (8)	S (6)
6	S (8)	VS (4)	ES (4)

- **A** What method was used to balance practice effects in this experiment?
- **B** Present the values you would use to describe the overall effect of the loudness variable. Include a verbal description of the effect along with the descriptive statistics that you use as a basis of your description.
- **C** What claim would you make about the effect of the loudness variable if the probability associated with the *F*-test for the effect of the loudness variable was $p = .04$?

Answer to Stretching Exercise

The first step is to compute a mean for each participant for each level of the independent variable by averaging responses across the two trials for the same condition. For Participant 1 the means for the three conditions are

10s	30s	50s
$(2 + 6)/2 = 4$	$(9 + 3)/2 = 6$	$(6 + 10)/2 = 8$

For Participant 2 the means for the three conditions are 3 (10s), 7 (30s), and 9 (50s), and for Participant 3 the means are 2 (10s), 5 (30s), and 7 (50s).

The next step is to compute the means for each condition by averaging the summary scores for each participant:

$$10s: (4 + 3 + 2)/3 = 3$$

$$30s: (6 + 7 + 5)/3 = 6$$

$$50s: (8 + 9 + 7)/3 = 8$$

We can now describe the effect of the independent variable, delay between target and test, on the dependent variable, number of errors. The means indicate that the number of errors on the pattern-identification task increased as the delay between target and test increased. Inferential statistics using null hypothesis testing or confidence intervals could be done to confirm whether the delay variable produced a reliable effect.

Answer to Challenge Question 1

A Assigning the values 1, 2, and 3 to the Low, Medium, and High conditions, respectively, and using the first row of the random number table (Table A.1) in the Appendix beginning with the first number in the row, the block-randomized sequence is Low-High-Medium-Low-Medium-High. One possible ABBA counterbalanced sequence is Low-Medium-High-High-Medium-Low.

B Because there are six conditions, all possible orders are not feasible. Therefore, either a Latin Square or a random starting order with rotation is needed to balance practice effects. A possible set of sequences using rotation is

	Position					
Participant	1st	2nd	3rd	4th	5th	6th
1	8	10	11	9	7	12
2	10	11	9	7	12	8
3	11	9	7	12	8	10
4	9	7	12	8	10	11
5	7	12	8	10	11	9
6	12	8	10	11	9	7