CHAPTER

8

SINGLE-FACTOR EXPERIMENTAL DESIGNS

CHAPTER OUTLINE

Can you unscramble the following anagram to form a five-letter word?

NIDRK

Imagine you are participating in a laboratory experiment. The experimenter explains that your task is to solve 15 five-letter anagrams, similar to the one above (which, unscrambled, spells *drink*). After you complete a 5-minute practice anagram test, the experimenter hands you an envelope that contains the real test. Like the practice test, this test has a cover page followed by five pages, each containing three anagrams. Unlike the practice test, on the real test your answers will be scored and you'll receive feedback. Therefore, the experimenter asks you to make sure that your test contains a participant ID code number, which has been pre-assigned to you and is handwritten in the upper-right corner of each page. You check, and indeed the code is there.

The experimenter leaves and you begin the task. Five minutes later, the experimenter returns, scores the test, and tells you the score. Next, you complete a questionnaire, the experimenter debriefs you, and your session ends.

Social-personality psychologist Andrew Elliot and his colleagues conducted this experiment (Elliot, Maier, Moller, Friedman, & Meinhardt, 2007). There's just one important detail I haven't revealed to you: Based on random assignment, the pre-entered ID code that appeared on your pages was written in either red, green, or black ink. And as the researchers predicted, overall, participants who were assigned red ID codes performed more poorly on the task than participants assigned green or black codes.

Many experiments, conducted by diverse researchers, have shown that the color of objects can influence people's behavior and psychological functioning (see Elliot, 2015). Yet, how can something seemingly as trivial as the color of a code number influence people's cognitive performance? Maybe the result was a fluke? Well, Elliot's research team replicated the experiment several times (Elliot et al., 2007; Elliot, Payen, Brisswalter, Cury, & Thayer, 2011). They studied American, German, and French undergraduates, and German high school students. They always included the color red, but varied the other colors (green or blue, and black, gray, or white). Instead of an anagram test, they used numerical, analogy, or memory tasks taken from IQ tests. And rather than manipulating the color of code numbers in the test booklet, the researchers varied the color of the test's cover sheet. In every replication of the experiment, participants performed worse overall when they were exposed to the color red. **Figure 8.1** presents the results from two of Elliot et al.'s (2007) experiments: one (a) examining how red, green, and black code numbers affected U.S. undergraduates' anagram performance, the other (b) how very brief exposure to red, green, or white on a test booklet cover sheet affected German undergraduates' performance on analogy items from an IQ test.

Recall from prior chapters that in an **experiment,** *the researcher manipulates one or more variables, attempts to control extraneous factors, and then measures how the manipulated variables affect participants' responses.* Recall as well that the term **independent variable** *refers to the variable manipulated by the researcher,* and the term **dependent variable** *refers to the response that is measured, to determine whether an independent variable has produced an effect.*

This chapter examines experiments that contain a single independent variable. In Chapter 9, we'll explore experimental designs containing two or more independent variables. In Chapter 10, we'll focus on issues of validity that pertain to experiments and potential pitfalls that experimenters seek to avoid.

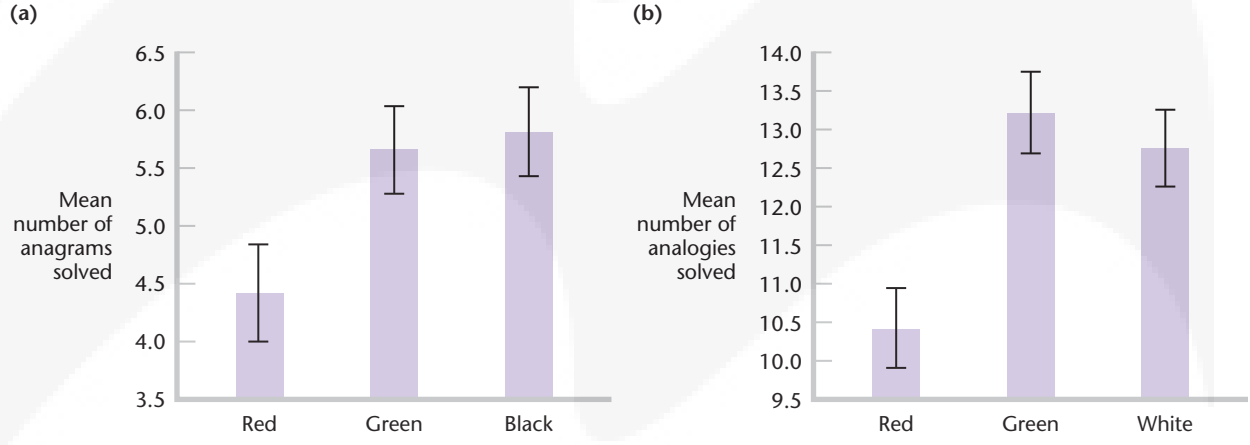


Figure 8.1 Effects of color on cognitive performance. **(a)** U.S. college students exposed to red ID codes correctly solved fewer anagrams than students exposed to green or black ID codes. **(b)** German undergraduates exposed to a partly red test booklet cover solved fewer analogy items than students exposed to a white or partly green cover. The thin I-shaped bars that straddle each large bar are called *error bars*. These particular error bars are called "95% confidence interval bars." They reflect the variability of scores within each condition. See Statistics Module 15 for more detail.
(Adapted with permission from the American Psychological Association from "Color and Psychological Functioning: The Effect of Red on Performance Attainment," by Elliot et al., 2007, *Journal of Experimental Psychology: General, 136*[1], p. 154.)

THE LOGIC OF EXPERIMENTATION

Learning Objectives

After studying this section, you should be able to:

- Describe the basic components of experimental control.
- Discuss how experimental control helps researchers to satisfy three key criteria for inferring cause and effect.

Psychologists conduct a dizzying array of experiments. One experimenter may disable a gene's functioning to determine whether it influences the feeding behavior of laboratory mice, while another may compare the effectiveness of different psychotherapies in treating depression. The goal of Andrew Elliot and his colleagues (2007, 2011) was to test the hypothesis that perceiving the color red would impair people's achievement on cognitive tasks. They based their hypothesis on learning theory: specifically, on classical conditioning principles (recall how Pavlov's dogs became conditioned to salivate to the sound of a tone). The researchers reasoned that, starting in childhood, the color red becomes associated with danger and poor achievement (e.g., red warning signs, teachers marking students' mistakes in red). Thus, the color red may become a conditioned stimulus that, without conscious awareness, increases people's sense of threat in achievement situations and thereby impairs their performance.

Regardless of the topic or whether a researcher's goal is to test a theory, all experiments share a common underlying method. You isolate a factor of interest (e.g., color), tinker around with it (e.g., expose people to different colors), try to keep other aspects of the situation constant, and then see what effects your tinkering produces. Although the experimental method has pitfalls and limitations, a well-crafted experiment is a scientific work of art: It is the best scientific tool available to draw clear conclusions about cause–effect relations. Let's examine why.

EXERCISING CONTROL OVER VARIABLES

In everyday life, one meaning of the word *control* reflects the ability to regulate or exercise direction over something. For scientists, the concept of **experimental control** *includes the ability to:*

1. *manipulate one or more independent variables;*
2. *choose the types of dependent variables that will be measured, and how and when they will be measured so that the effects of the independent variables can be assessed; and*
3. *regulate other aspects of the research environment, including the manner in which participants are exposed to the various conditions in the experiment.*

The goal of having such control is to enable researchers to conclude that the variable they manipulate—rather than other uncontrolled factors—is the cause of any obtained effects on behavior. For example, in their initial experiment, Elliot et al. (2007)

- manipulated the color of the ID code that was placed on the test;
- chose to measure cognitive performance via an anagram task, and controlled when the task was given;
- regulated other aspects of the research setting (e.g., task instructions were always the same), including how participants were exposed to the various colors (i.e., they were randomly assigned).

By carefully exercising these types of control, the researchers were able to conclude that the variation in ink color—rather than some other aspect of the situation—was the most plausible explanation for why participants' performance differed across the three color conditions.

CAUSAL INFERENCE AND EXPERIMENTAL CONTROL

How does experimental control enhance our ability to draw causal conclusions? Recall that three criteria need to be met in order to conclude that variable X has a causal influence on variable Y:

1. *Covariation of X and Y.* As X varies, Y varies.
2. *Temporal order.* The variation in X occurs before the variation in Y.
3. *Absence of plausible alternative explanations.*

In principle, if X is the only factor in a situation that varies prior to a change in Y, then the logical conclusion is that the variation in X must have caused the change in Y.

Now let's map the different components of experimental control onto these three causal criteria. The first causal criterion, the covariation of X and Y, is achieved by manipulating the independent variable—by creating two or more distinct conditions in an experiment—and measuring whether scores on the dependent variable differ among those conditions. Elliot et al. (2007) manipulated the independent variable, color, by creating red, green, and black ID code color conditions. They measured the dependent variable, cognitive performance, by determining the average number of anagrams solved by participants in each color condition. Their data analysis indicated that X and Y did covary: The number of anagrams solved differed significantly depending on the color condition.

The second causal criterion, temporal order, is achieved by exposing participants to the manipulated independent variable prior to any changes that may occur in the dependent variable. In the color experiment, participants were first exposed to one of three colors and then performed the anagram task. There is no way that Y (anagram performance) could have caused X (color exposure).

Researchers attempt to achieve the third causal criterion, the absence of plausible alternative explanations, by eliminating other factors that might systematically cause Y to vary. For example, there are many variables that might influence people's performance on an anagram task, such as how hot, humid, and noisy the room is while performing. Although some researchers might be interested in studying how these factors affect performance, Elliot et al. (2007) were not; they were interested in how color affects performance. Thus, to Elliot and his colleagues, each of these other factors represented an **extraneous variable:** *a factor that is not the focus of interest in a particular study, but that could influence the outcome of the study if left uncontrolled.* Therefore, in conducting their experiment, Elliot et al. would want to keep these extraneous environmental factors as constant as possible.

Let's consider two other extraneous variables. First, suppose that in the color experiment, the red ID code had been very large in size, whereas the green and black codes were small. In this case, perhaps it wasn't the code's red color that caused participants to perform more poorly, but its larger size; maybe a large-sized ID code of any color would have distracted participants. Second, suppose that the college students in the red, green, and black conditions participated in different weeks of the study (i.e., weeks 1, 2, and 3, respectively). Imagine that week 1 (when the red-condition students participated) was the week of midterms, and weeks 2 and 3 (the green-condition and black-condition participants) occurred after midterm week. Might it be reasonable to speculate that anxiety, distraction, or fatigue surrounding midterm exams caused the students in the red color condition to perform more poorly than the students in the other color conditions?

In these cases, the size of the ID code and the proximity to midterm exam week would represent extraneous factors that have now become confounding variables. A **confounding variable** *is a factor that covaries with the independent variable in such a way that we can no longer determine which one has caused the changes in the dependent variable.* **Table 8.1** illustrates how our hypothetical confounding variables covary with the independent variable of color.

Table 8.1 Potential Confounding of an Independent Variable

	Condition 1	Condition 2	Condition 3
Independent Variable Color (of ID code)	Red	Green	Black
Confounding Variable 1 Size of ID code on page	Large	Small	Small
Confounding Variable 2 Proximity to midterm exams	Midterm week	1 week after midterms	2 weeks after midterms

Good researchers acquire the ability to recognize ahead of time potential confounding variables that could ruin the interpretability of their studies. Researchers then design their studies to eliminate or at least minimize those confounding variables. In experiments, researchers typically reduce confounding variables by:

- keeping extraneous factors as constant as possible across the different conditions of the experiment; and
- balancing extraneous factors that, in principle, cannot be held constant.

Let's examine how these procedures are applied to potential confounding variables that arise from two sources: the research environment and characteristics of the research participants.

Potential Confounding Variables: Environmental Factors

Many extraneous environmental factors that might confound the results of an experiment can, in principle, be held constant or at least nearly so. All participants can perform the same anagram task in the same room and sit in the same chair. The room's temperature, humidity, and noise level can be kept reasonably steady.

There are other environmental factors that even in principle cannot be held constant, but that can be balanced across the different experimental conditions. These include the time of day, day of the week, and month of the year when each student participates. Thus, on any given day, as the researcher, I would try to "run" (i.e., conduct) the experiment with an equal number of participants from each condition in the morning and in the afternoon. This way, after the experiment is finished, you would not be able to say, "I think the reason that performance was worst in the red color condition was that those participants always performed the anagram task on Friday afternoon, at the end of the tiring school week, or early on Monday mornings while still recovering from a weekend of partying, whereas students in the other conditions always performed the task midweek." Over the entire experiment, these factors—time of day, day of week—will end up being as equivalent as possible across the three color conditions.

The experimenter who interacts with the participants also constitutes a part of the environment, and would be trained to behave in a standardized manner. However, the experimenter isn't a robot, and her or his behavior will fluctuate at least minimally. What researchers hope to avoid is a *systematic bias* in which the experimenter consistently alters her or his behavior toward participants based on the particular condition they are in.

Potential Confounding Variables: Participant Characteristics

No matter how successfully researchers have tamed potential environmental confounding factors, there is still one possible objection to the conclusion that, for example, the red color of the ID code caused poorer anagram performance. Perhaps the students in the red color

condition would have performed the most poorly anyway, even if they had been shown a green or black ID code. Maybe, as a group, they simply had poorer anagram ability, less confidence, or lower levels of intelligence than the participants in the other color conditions. If so, their poorer performance might have had nothing to do with the color red.

Potential confounding variables due to participants' characteristics are addressed differently in two major approaches to designing experiments. With a **between-subjects design,** *different participants are assigned to each of the conditions in the experiment*. In this design, researchers typically minimize the potential confounding effects of subject characteristics by using **random assignment,** *a procedure in which each participant has an equal probability of being assigned to any one of the conditions in the experiment*. Random assignment distributes participants' individual differences across the experimental conditions in an unbiased, nonsystematic way. Thus, at the start of the color experiment, the three groups of participants are assumed to be equivalent to one another.

In a **within-subjects design,** *each participant engages in every condition of the experiment one or more times*. Because the same people participate in all the conditions, relatively stable factors such as participants' general anagram ability and intelligence essentially remain the same in each condition and cannot confound the results. However, factors such as boredom, fatigue, and experience with the task can change as each participant moves from one condition of the experiment to the next. If every participant engages in all the conditions in the same order with, say, the color red always being the last condition, we have no way of knowing whether the red color caused the poorer performance or whether participants might have been fatigued toward the end of the experiment. The solution to this problem is **counterbalancing,** *a procedure in which the order of conditions is varied so that no condition has an overall advantage relative to the other conditions*. Before we explore between- and within-subjects designs more closely, in the next section we'll discuss how independent variables can be manipulated.

✓ CONCEPT CHECK 8.1 THE LOGIC OF EXPERIMENTATION

Match each concept on the left with its correct description on the right. Answers appear at the end of the chapter on page 276.

1. between-subjects design	**(a)** a factor manipulated by the experimenter
2. within-subjects design	**(b)** examples are random assignment, counterbalancing
3. experimental control	**(c)** each participant engages in only one condition
4. dependent variable	**(d)** each participant engages in all conditions
5. confounding variable	**(e)** provides an alternative explanation for the findings
6. independent variable	**(f)** an outcome measured by the experimenter

MANIPULATING INDEPENDENT VARIABLES

Learning Objectives

After studying this section, you should be able to:

- Describe ways to manipulate an independent variable.
- Discuss factors that affect the number of conditions incorporated into an experiment.
- Explain the concepts of experimental and control groups.

In designing an experiment, we create an independent variable by forming two or more conditions that vary the amount or type of some factor. In other words, we can manipulate independent variables quantitatively and qualitatively.

VARYING THE AMOUNT OR TYPE OF A FACTOR

As people consume greater quantities of alcohol at a party, how does this affect their behavior? Does the size of a group influence the likelihood that a person will conform to the group's opinions? To study each question experimentally, we would manipulate the amount of alcohol we ask people to consume and the size of the group to which we expose people. These are quantitative manipulations and we could label these independent variables "alcohol dose" and "group size," respectively.

In contrast, consider two other questions. First, what type of psychotherapy most effectively treats moderate depression? Second, does our ability to accurately recognize emotions from facial expressions depend on the type of emotion? To study these questions experimentally, we would create different conditions, each of which represents a different type of psychotherapy or emotional expression. These are qualitative manipulations and would produce independent variables that we could label "type of therapy" and "type of emotion," respectively. In Elliot et al.'s (2007, 2011) color experiments, the independent variable was manipulated qualitatively: Participants were exposed to one of three colors.

As for the variety of factors that experimenters manipulate, here are some common approaches to creating independent variables (note that some may overlap):

- *altering the physical environment,* such as exposing people to different room temperatures, intensities or types of sounds, or schedules or types of reinforcement during a learning task;
- *altering the social environment,* such as leading participants to believe that a job applicant is female or male, exposing infants to familiar or unfamiliar voices, or varying the size of groups;
- *varying an intervention provided to people,* such as the type or amount of psychotherapy, counseling, or skills training;
- *varying the task that participants perform,* such as asking people to remember lists of words of varying lengths, or abstract versus concrete words;
- *varying the strategy that people are instructed to use when performing a task,* such as instructing people to learn material either by rote memorization or by using visual imagery;
- *manipulating an organism's characteristics,* such as in laboratory mice, disabling a particular gene or surgically lesioning a specific brain area to see how behavior is affected; or manipulating the type or intensity of people's mood by exposing them to pleasant or unpleasant stimuli, to determine how mood influences their subsequent task performance.

DETERMINING THE NUMBER OF CONDITIONS

In experiments, a **single-factor design** *has only one independent variable.* This independent variable must have at least two *conditions,* also called two *levels* of the independent variable. An experiment with one independent variable that has more than two levels is often called a *single-factor, multilevel design.* **Table 8.2** provides examples of how independent variables can be manipulated so they have more than two levels.

The Research Question and Available Resources

What determines how many levels of an independent variable we should create? Essentially, the question about behavior that we ask, our personal preferences, and our assessment of

Table 8.2 Creating Independent Variables with Two or More Levels

Independent Variable	Design	
	Two-Level	Multilevel
Sleep deprivation	Total deprivation (8 hours) No deprivation (0 hours)	Total deprivation (8 hours) Partial deprivation (4 hours) No deprivation (0 hours)
Length of word list	15 words (long list) 3 words (short list)	15, 12, 9, 6, or 3 words
Monetary incentive	Low incentive High incentive	No incentive Low incentive Moderate incentive High incentive
Described sex of hypothetical job applicant	Female Male	Female Male No description (sex not mentioned)
Type of therapy	Psychodynamic therapy Cognitive therapy	No therapy Psychodynamic therapy Cognitive therapy Behavior therapy

available resources will be three key factors. If we are interested in age stereotyping, then we could ask participants to evaluate a job résumé, and vary whether we tell participants that the job applicant is 30 or 50 years old. If instead our research interest is ethnic stereotyping, then although we could limit the independent variable (i.e., ethnicity of job applicant) to two conditions, creating more conditions would yield more information. Thus, we could lead participants to believe that a 30-year-old female applicant, born in Chicago, is of African, Asian, Hispanic, White Northern European, or Native American descent.

As another example, if our research question is "How does fully depriving people of a night's sleep affect them psychologically?," then creating an experiment with two deprivation levels—8 hours (i.e., "total sleep deprivation") and 0 hours (i.e., "normal sleep control group")—may be sufficient. But, if we want to know how different degrees of sleep deprivation affect people, then we'll want to create more conditions. For example, one research team deprived people of 0 hours, 2 hours, 4 hours, or 8 hours of sleep and then measured how this affected participants' daytime sleepiness and cognitive performance (Roehrs, Burduvali, Bonahoom, Drake, & Roth, 2003).

Balanced against our desire to create more levels of an independent variable, we must consider practical issues, such as the resources available to us. Will enough participants be available? If they're to be paid or reimbursed for expenses, do we have sufficient funds? Is the sleep laboratory available for all the hours we need? Sometimes a researcher's first experiment on a topic may contain only one independent variable and just two or three conditions. Then, to further explore the topic, the researcher designs subsequent experiments that have more independent variables and/or more levels of an independent variable.

The Potential to Examine Nonlinear Effects

In Solomon Asch's (1955) classic experiments on conformity, groups of varying sizes performed a series of simple visual tasks: judging the length of lines. Only one member of each group was a real participant, and this participant was unaware that the other members were confederates who, according to plan, voiced clearly incorrect judgments on some of the trials. Asch manipulated group size by controlling the number of confederates in each group, and

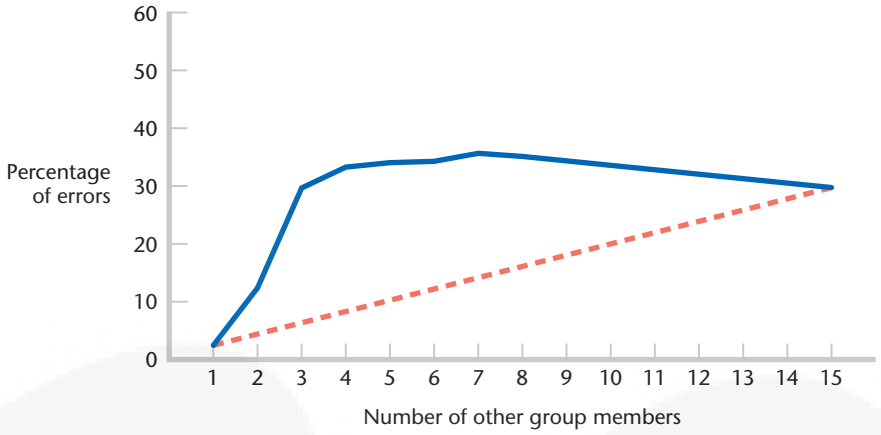


Figure 8.2 Relation between group size and conformity as determined by Solomon Asch. Asch found a nonlinear relation between group size and conformity to the group (see the solid line). Had he only included two group sizes in his experiment (see the dashed line), the nonlinear relation would not have been detected.

he measured the percentage of trials on which the real participants conformed to the group's erroneous opinion. As group size increased, would conformity increase in a linear fashion? As the solid line in **Figure 8.2** shows, Asch's finding was "no": Conformity rose as group size increased from about 1 to 4 other members, but after that, further increases in group size had little effect on conformity.

Looking at the dashed line in Figure 8.2, imagine that Asch (1955) had manipulated group size by creating only two conditions: groups with 1 other member, and with 15 other members. Would this straight line accurately portray the relation that Asch actually found? Of course not. If we quantitatively manipulate an independent variable by creating just two conditions, then any line graph of the data can only portray a straight line. Thus, there is no way to know whether this reflects an actual linear relation between the independent and dependent variables, or instead is merely an artifact imposed by the limitation of having only two data points in the graph. To examine whether a nonlinear relation exists—and to provide a more thorough test of a potential linear relation—we need to create an independent variable that has three or more levels.

As another example, the solid line in **Figure 8.3** illustrates a nonlinear relation between the amount of alcohol that young adults were administered in an experiment, and their perception of how much that alcohol impaired their cognitive task performance (Roehrs et al., 2003). Scores further below or above the *y*-axis point of 0 indicate greater amounts of perceived performance impairment and performance enhancement, respectively.

In this study, experimental-physiological psychologist Timothy Roehrs and his colleagues manipulated the dose of alcohol by having participants drink 0.0., 0.3, 0.6, or 0.9 grams of alcohol per kilogram of body weight (Roehrs et al., 2003). This produced, on average, breath alcohol concentrations of .00, .02, .04, and .09 a half hour after consumption. The graph shows that it was not until the highest dose of alcohol was administered that participants, overall, believed their performance was impaired. Had the researchers only used no alcohol and high dose conditions, as illustrated hypothetically by the dashed line, this would have masked the nonlinear relation.

What makes this multilevel design especially interesting is that Roehrs et al. (2003) also measured participants' daytime sleepiness and actual cognitive performance. Alcohol increased sleepiness and decreased performance in a linear fashion. Given these linear

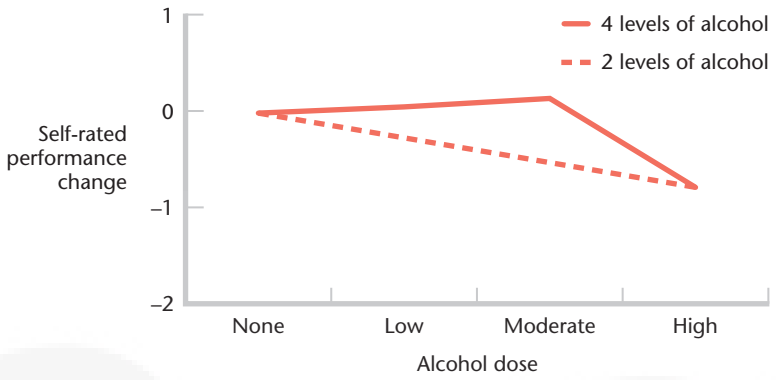


Figure 8.3 Self-perceived performance impairment as a function of alcohol dose. By manipulating four alcohol dosage levels, Timothy Roehrs and colleagues were able to examine whether the relation between alcohol dose and self-perceived performance impairment was nonlinear (solid line). Had only two dosage levels been manipulated (dashed line), the nonlinear relation would not have been detected. (Reprinted/adapted from Roehrs et al., *SLEEP* 2003; 26:981-5 with permission from Associated Professional Sleep Societies, LLC, Darien IL, 2016.)

relations, but the nonlinear relation between alcohol dose and perceived impairment, the researchers concluded that participants failed to recognize alcohol's impairing effects until they had consumed enough alcohol, on average, to be legally drunk in most jurisdictions.

EXPERIMENTAL AND CONTROL CONDITIONS

In many experiments, comparisons are made between experimental conditions and control conditions. An **experimental condition** (or **experimental group**) *involves exposing participants to a treatment or an "active" level of the independent variable*. In a **control condition** (or **control group**), *participants do not receive the treatment of interest or are exposed to a baseline level of an independent variable*. Depending on how an experiment is designed, different participants may be assigned to the experimental and control conditions, or the same participants may serve in both conditions.

One common way to create experimental and control conditions is to manipulate the *presence versus absence* of some factor. Suppose we want to examine whether exposure to traffic noise affects people's cognitive task performance. We expose an experimental group to an "active" level of the independent variable: Traffic noise will be present. The control group is not exposed to noise and this presents the baseline condition against which performance in the noise condition will be compared.

Instead of noise, to examine how high temperature affects performance, we can have participants in an experimental condition perform a task in a 100 °F (37.8 °C) room. This would represent the active level of the independent variable, but there's no such thing as creating a control condition in which an "absence" of temperature exists. Every room has a temperature, and exposing people to a frigid temperature of 0.0 °F (or 0.0 °C) would hardly be a "control"! So, we might create our control condition—our standard of comparison—by having a group perform the task at 72 °F (22.2 °C), a room temperature most people find comfortable.

There are also experiments in which the concept of a control group does not apply. Suppose we conduct an experiment to determine how accurately people can identify facial expressions of six emotions: anger, fear, joy, disgust, surprise, and sadness. Participants view a series of 60 photographs, created by having each of 10 actors portray the six emotions. In this study, each of the six emotions represents an experimental condition and there is no control condition per se (i.e., the design does not include a "no emotion" condition). However, if

we were interested in examining the question "Do people recognize each of these emotional expressions more accurately than a neutral facial expression?," then we would need to add a control condition by including photos in which the same 10 actors each portray a neutral facial expression. Thus, whether we design our emotion recognition experiment to include a control group depends on the question we are examining.

Experiments sometimes call for special control groups to address potential confounding variables that a traditional "no treatment" or "baseline exposure level" control group cannot eliminate. We'll discuss these control groups in Chapter 10.

CONCEPT CHECK 8.2 MANIPULATING INDEPENDENT VARIABLES

Decide whether each statement below is true or false. Answers appear on page 276.

1. Experimenters can only manipulate environmental factors; they cannot manipulate an organism's characteristics.
2. Independent variables must be quantitative; they cannot be qualitative.
3. A single-factor experimental design must include at least two conditions.
4. To examine nonlinear effects, researchers should use a multilevel design.
5. Some experiments do not include a control group.

BETWEEN-SUBJECTS DESIGNS

Learning Objectives

After studying this section, you should be able to:

- Explain the advantages and disadvantages of between-subjects designs.
- Describe several types of between-subjects designs.
- Discuss the differences between random assignment and random sampling.

Typically, in single-factor experiments, each participant either engages in only one condition, or engages in all the conditions. As noted earlier, these two approaches are called, respectively, a *between-subjects design* and a *within-subjects design*. Suppose we want to examine whether taking class notes longhand—as compared to taking class notes using a laptop computer—enhances, impairs, or has no significant influence on college students' learning of lecture content. The left side of **Figure 8.4** shows how we could investigate this issue by conducting an experiment that employs a basic between-subjects design. This design is basic in the sense that it only has two conditions, the minimum needed to conduct a between-subjects experiment. In one condition, students take longhand notes during a lecture, and in the other condition, students use a laptop to take notes during the same lecture. Each student in this experiment would participate in only one of these conditions.

For visual simplicity, the left half of Figure 8.4 shows a total of 20 participants in the experiment (in reality, we would want a larger sample size). Based on random assignment, 10 students are assigned to take longhand lecture notes and 10 others are assigned to use a laptop computer. To guarantee that each student is exposed to the same lecture content, delivery style by the lecturer, and note-taking environment, we could video-record a lecture and have each student view it in a laboratory, individually, so as not to be influenced by the presence and potential note-taking behavior of other students. Afterward, we would administer the same knowledge test (e.g., a multiple choice exam) to assess each student's understanding of the lecture material.

Pam Mueller and Daniel Oppenheimer (2014) compared the effectiveness of longhand versus laptop note-taking by conducting a series of experiments. Although their procedures varied from one experiment to the next, the key point is that they manipulated longhand versus laptop note-taking as a between-subjects independent variable in each experiment.

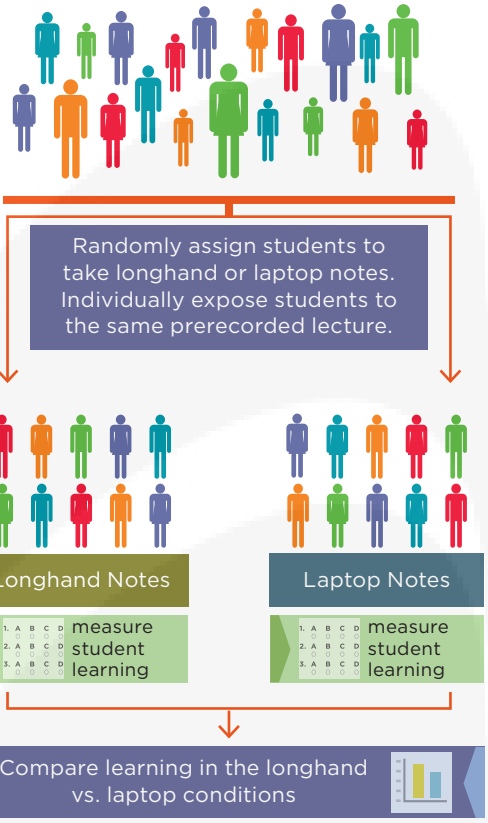
Infographic: **Figure 8.4**

BETWEEN-SUBJECTS AND WITHIN-SUBJECTS EXPERIMENTAL DESIGNS

Research Question: Compared to taking longhand lecture notes, does laptop note-taking enhance or impair students' learning?

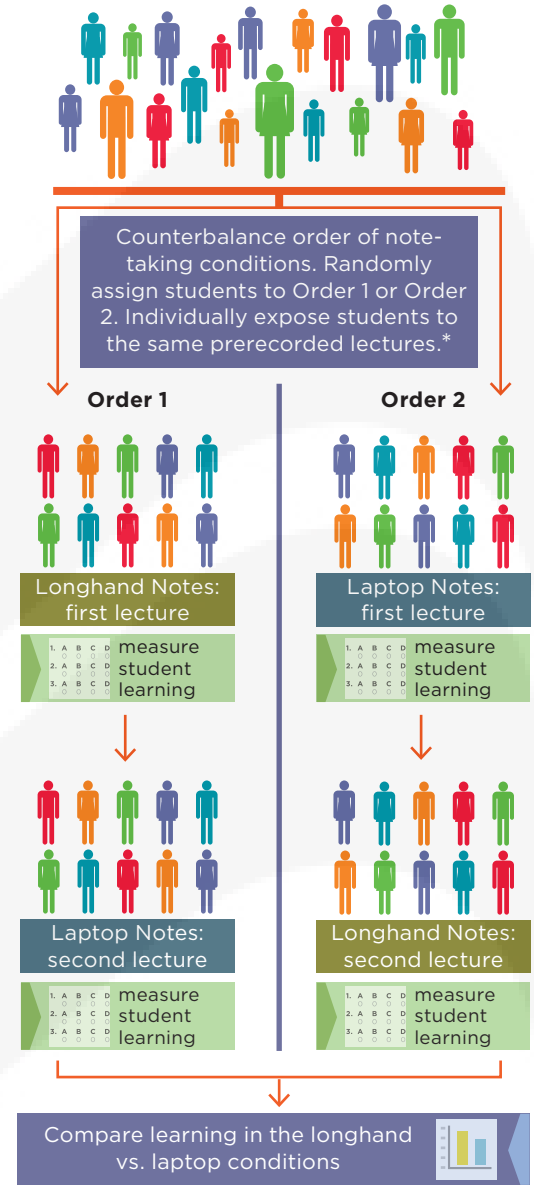
Between-Subjects Design

(Each participant engages in only one condition.)



Within-Subjects Design

(Each participant engages in every condition.)



*In the within-subjects design, it would not make sense to expose students to the same lecture and knowledge test twice. So we need two lectures and must decide on their order of presentation. How would you do this? We will discuss this on pages 264–265.

Using a similar between-subjects design, Mueller and Oppenheimer (2014) found that, compared to taking longhand notes, using laptops led students to take more notes, but also more verbatim notes, as opposed to writing notes in their own words. Laptop note-takers also showed poorer conceptual learning of the lecture content. The researchers concluded that laptop use promoted shallower, verbatim note-taking, which impaired students' learning of concepts.

Figure 8.4 describes their general findings and also illustrates how this experiment could be conducted using a within-subjects design, as we will discuss shortly.

ADVANTAGES OF BETWEEN-SUBJECTS DESIGNS

Between-subjects designs offer several advantages. First, because each participant only engages in one condition, effects caused by exposure to one condition (e.g., stress, fatigue) can't carry over to other conditions. Second, some scientific questions can only be examined using a between-subjects design. Consider a classic experiment by comparative psychologist Eckhard Hess (1959) on *imprinting,* a biologically primed form of attachment that occurs in some bird and mammalian species. If you've ever seen an adult duck or goose waddle or swim from one point to another, while being followed by a procession of its young, you've witnessed the end result of imprinting.

Hess (1959) hatched and raised mallard ducklings in a laboratory, isolating each one. He then manipulated the number of hours after hatching when each duckling was briefly removed from isolation and exposed to a realistic wooden model of an adult mallard duck. The wooden model was mechanically moved around a circular runway for 10 minutes, and the duckling could follow the model by waddling around a circular path (**Figure 8.5**). Later, Hess tested whether the duckling had imprinted on the model by measuring whether it followed either the original model or a different model duck it was seeing for the first time.

In his study, Hess (1959) found that imprinting was most likely to occur when a duckling was exposed to the model between 13 and 16 hours after hatching; by 32 hours, imprinting was successful for only a small percentage of the ducklings. Hess thus concluded that there was a "critical period" in imprinting, a particular age range during which exposure to the parent must occur or the capacity to imprint will be lost. In this experiment, a between-subjects design had to be used to manipulate the ducklings' age of initial exposure, because just as you only get one chance to make a first impression, there is only one opportunity to give each duckling its first exposure to the adult model.

A third advantage of between-subjects designs is that, even when it's possible to expose participants to every condition, doing so may require developing different but equivalent versions of the same task, and it also may tip off participants about the hypothesis being tested or true purpose of a study. This additional complexity may be avoided by using a between-subjects design. For example, in our between-subjects note-taking experiment, we can expose all students to the same lecture, and thus have to develop only one lecture to serve as the stimulus material. In contrast, using a within-subjects design, it would not make sense to have students view the same lecture twice and take the identical knowledge test, once when taking longhand notes and again when taking laptop notes. Instead, we would need to develop two equivalent lectures and knowledge tests, as we'll discuss more fully when covering within-subjects designs.

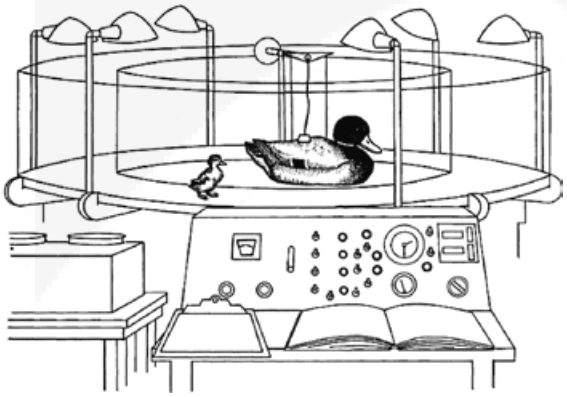


Figure 8.5 Drawing of the imprinting apparatus used by comparative psychologist Eckhard Hess. (Hess, 1959.) (Reprinted with permission of *Science* from "Imprinting," by E. H. Hess, 1959, *Science*. Permission conveyed through Copyright Clearance Center, Inc.)

As another example, consider the color experiment by Elliot et al. (2007) described at the beginning of the chapter. Obviously, if each student were to participate in each of the three color conditions, we could not give them the same anagrams to solve each time! Thus, we would need to develop three sets of anagrams of equal difficulty. By using a between-subjects design, Elliot et al. were able to expose all participants to the same, single set of anagrams. In addition, if you were a participant, after first receiving an

anagram packet with, say, a red ID code, and then one with a green code, and then a third one with a black code, might you begin to suspect that the experiment had something to do with color? And once this happened, you might alter your behavior. In a between-subjects design, participants are less likely to become sensitized to the hypothesis being tested because they are not exposed to—and therefore are unlikely to become aware of— the different manipulations of the independent variable that constitute the various conditions in the experiment. Based on discussions held with participants during debriefing sessions, Elliot et al. found that not a single participant in any of their experiments suspected the true purpose of the study.

Finally, there may be other reasons why it would be inadvisable to have each individual participate in all the conditions. For example, a task may be too time consuming, too fatiguing, or too stressful to either practically or ethically have people perform it in each condition of the experiment.

DISADVANTAGES OF BETWEEN-SUBJECTS DESIGNS

Of course, between-subjects designs have disadvantages as well as advantages. One disadvantage is that, although they usually are effective in creating equivalent groups across the various conditions of an experiment, they are less effective at this than are within-subjects designs. Simply put, in trying to keep extraneous factors such as participants' backgrounds, personality traits, and abilities as constant as possible across the various conditions of an experiment, what could be more equivalent than having the same participants engage in every condition?

With a between-subjects design, if we find a difference between how an experimental group and a control group have responded on the dependent variable, some portion of that difference may result not from the different levels of the independent variable itself, but from overall differences (even if small) between the groups in their preexisting personal characteristics. Therefore, in a between-subjects rather than a within-subjects design, the independent variable is essentially going to have to "work harder" (i.e., it will need to produce a bigger effect) to be noticed. Stated another way, if an independent variable truly has an influence on behavior, then overall, between-subjects designs will be less sensitive to detecting that effect than will within-subjects designs.

Another disadvantage of between-subjects designs is that the experimenter has to recruit and study more participants—often many more—in order to gather the same amount of information. Notice in Figure 8.4 that, using a between-subjects design, we end up with data from 10 students assigned to the longhand condition, and from 10 students assigned to the laptop condition. By employing a within-subjects design, we would end up with data from 20 students in each condition, because each student would participate in both conditions. The disadvantage of between-subjects designs in terms of recruiting participants becomes more prominent as the number of conditions in an experiment and the desired sample size per condition increase. For example, in an experiment with four conditions, if we want to gather data from 30 participants per condition, we'll need to recruit 120 participants in a between-subjects design, but only 30 using a within-subjects design.

TYPES OF BETWEEN-SUBJECTS DESIGNS

Now, let's examine two between-subjects experimental designs. We'll also discuss whether another common between-subjects design represents an experimental design.

Independent-Groups Design

In an **independent-groups design** (also called a **random-groups design**), *participants are randomly assigned to the various conditions of the experiment*. For example, in the color

experiment, each participant was exposed to only one color—red or green or black—and that determination was made randomly. Random assignment means that each participant has an equal probability of being assigned to any particular condition.

Random assignment does not eliminate the fact that each research participant has a unique psychological and biological makeup, a unique set of experiences, and so forth. Nor does random assignment eliminate the reality that these individual differences may influence how people respond during the experiment. Rather, random assignment distributes these individual differences across the various conditions in an unbiased way. This enables researchers to assume with a high level of confidence that overall, prior to exposure to the independent variable, the groups of participants in the various conditions are equivalent to one another (1) on the attributes that will be assessed by the dependent measures, and (2) on other personal characteristics that might influence responses to the dependent measures.

Note that "equivalent" does not mean that the groups in the various conditions will be identical or perfectly matched on their personal characteristics at the outset of the experiment. Rather, it means that such preexisting differences will likely be too inconsequential to provide a plausible alternative explanation for any statistically significant findings that our experiment may obtain.

For example, in the color experiment, we can be highly confident in assuming that at the outset, participants in the red, green, and black conditions were equivalent, overall, in their ability to solve anagrams. Moreover, if factors such as intelligence and past experience with anagrams might influence current anagram performance, by using random assignment, we can be confident that the participants in the three color conditions were equivalent overall on these characteristics as well. Therefore, when participants performed most poorly after being exposed to the color red, we can confidently discard the alternative explanation that this was caused by poorer general anagram ability, lower intelligence, or less prior anagram experience.

You may have noticed that I've used terms such as *highly confident* rather than *absolutely sure*. Random assignment cannot guarantee the equivalence of groups with absolute certainty. There is always a possibility that purely by chance, participants will be distributed across the various conditions in a way that does create an important preexisting overall difference in their general anagram ability, intelligence, prior experience, and so forth. This can happen because, with chance, once in a while improbable outcomes do happen. To provide a rough analogy, if we flip a coin 10 times, the odds are less than 1 in a 1,000 that we'll get a Heads each time, but there is no guarantee that it won't happen.

This example brings up a related point: The greater the number of participants, the more likely random assignment will successfully create equivalent groups. Flip a coin only twice, and the odds of getting an extreme outcome of all Heads are fairly high: 25% (i.e., HH, HT, TH, TT). Flip it 20 times, and the odds of obtaining all Heads become infinitesimal, and the odds of getting any outcome with a large imbalance of Heads and Tails also drop dramatically. Likewise, as random assignment is used with a larger number of participants, the laws of chance have a greater opportunity to even things out. It becomes increasingly unlikely, for example, that most of the participants with poorer anagram ability would happen to end up in the "red condition."

Note that researchers typically don't flip coins or draw straws to randomly assign each participant. Rather, an experimenter would use a *random number table* (a table of randomly ordered sequences of digits) or a random number generator (a software program), many of which can be accessed online. Table 6 in Appendix C contains a portion of a random number table.

Block Randomization. If you flip a coin repeatedly, by chance you will get sequences of several Heads or several Tails in a row. Likewise, with basic random assignment, you will get sequences in which several participants in a row will be assigned to the same condition.

This is not ideal. Imagine, for example, that by chance several sequences result in too many participants being assigned to the red condition on a Friday afternoon or Monday morning. With random assignment, it's also likely that you'll end up with a slightly different number of participants in each condition (e.g., flip a coin 20 times, and the odds of getting precisely 10 Heads and 10 Tails are only 17.6%). Fortunately, there's a simple procedure that addresses both of these issues.

Using **block randomization,** *we conduct a single round of all the conditions, then another round, then another, for as many rounds as needed to complete the experiment. Within each round, the order of conditions is randomly determined*. Here's an example, based on Elliot et al.'s (2007) first color experiment. We have three conditions—red, green, and black—and we want 30 participants per condition. The experiment will begin next Monday, with three morning sessions and three afternoon sessions. Each session involves one participant. For Monday morning, we randomly order the three conditions, say, black–red–green. This constitutes the first "block" (i.e., "set" or "round") of conditions. Our 9 a.m. participant is assigned to receive the black ink code, the 10 a.m. participant the red ink code, and the 11 a.m. participant the green ink code.

For Monday afternoon, we randomly order the three conditions again, say, green–black–red. This is our second block of conditions. Respectively, the 1 p.m., 2 p.m., and 3 p.m. participants are assigned to the green, black, and red ink codes. We use this procedure each day, randomly ordering the three conditions for each new block. Because we want 30 participants per condition, we will conduct a total of 30 blocks. Following this approach, by the end of the experiment, each color condition ideally will have been conducted not only an equal total number of times, but also an equal number of times in the morning, in the afternoon, on each weekday, and during each week.

Matched-Groups Design

Instead of assigning participants on a purely random basis, let's now consider another approach to creating equivalent groups in a color-achievement experiment. We might try to match the groups in the three color conditions on their general anagram ability. Perhaps we could give all the participants an anagram test at an earlier point in time. Next, we would take the three participants with the highest anagram ability and randomly assign one of them to each color condition. Then we would take the next three best performers and randomly assign one of them to each condition. We would continue to do this until we reached the three poorest performers, and likewise, we'd randomly assign one of them to each color condition.

What we've done is to use anagram ability as a **matching variable,** *a characteristic on which we match sets of individuals as closely as possible*. Then, to create a **matched-groups design,** *each set of participants that has been matched on one or more attributes is randomly assigned to the various conditions of the experiment*. If our experiment had only two conditions—say, red and black—then we would match pairs of participants and randomly assign a member of each pair to each condition. This is often called a *matched-pairs design*.

It might seem that experimenters should always match their participants on the same attribute that will be assessed later by the dependent measure, but it is often impossible or impractical to do this. Moreover, in some cases, giving participants the same task or type of task that they will subsequently perform might tip them off about the purpose of the experiment.

Instead of matching participants on the same variable that constitutes the dependent measure, an alternative is to match them on some other variable that we are concerned about as a possible confounding factor. For example, if we know that intelligence and verbal aptitude are substantially correlated with anagram performance, then perhaps we would want to try using these as matching variables rather than leave the creation of equivalent groups to

simple random assignment. This way, if the red color group performs worst, and if this group overall is matched with the other groups on IQ and verbal aptitude, then it's difficult to argue that IQ and verbal aptitude—rather than the red color—were responsible for our results.

To match the participants according to these criteria, we would either have to administer IQ and verbal aptitude measures to them, or try to obtain records of existing IQ or verbal aptitude scores. Sometimes, for reasons of confidentiality or other factors, it's not feasible to create correlated matching variables, but other times it is. For example, in an experiment in which two groups of elementary schoolchildren will be randomly assigned to receive one of two types of special tutoring programs, we might want to match the two groups on readily measurable characteristics such as gender and school grade.

Even when matching is feasible, however, the benefits need to be weighed against the cost of adding complexity to the experiment. By obtaining a sufficiently large number of participants for their experiments, many researchers choose to rely solely on random assignment to create equivalent groups and let chance "balance out" what otherwise might be too many potential matching variables to keep track of.

Natural-Groups Design

In many studies, researchers who are interested in examining a possible causal relation between variables X and Y create the different conditions of X not by manipulating an independent variable, but instead by selecting different groups of people based on their personal characteristics. Consider this question: Does self-esteem influence whether people are introverted or extraverted?

To answer this question, we design a study with high-self-esteem and low-self-esteem conditions. To create these conditions, we administer a psychological test that measures self-esteem to 200 college students. We identify the 40 students with the highest self-esteem, the 40 with the lowest self-esteem, and these groups operationally define our high- and low-self-esteem conditions. We then have each of these students individually come to our laboratory. In the lab, we expose the students to a situation in which they have the opportunity to interact with other people, and we record the degree to which their behavior is introverted or extraverted.

Subject Variables. People's natural level of self-esteem is a **subject variable:** *a personal characteristic on which individuals vary from one another.* What we have done in our self-esteem example is to develop a single-factor design by measuring (not manipulating) a subject variable and creating different conditions (i.e., groups) based on people's natural level of that variable. This research approach is called a **natural-groups design:** *A researcher measures a subject variable, forms different groups based on people's level of that variable, and then measures how the different groups respond on other variables.*

Similarly, depending on our research question, we can use other subject variables to create a natural-groups design. If we hypothesize that people's personality and values differ depending on their degree of morningness–eveningness (the tendency to be an "early bird" or "night owl"), then we could measure morningness–eveningness and create a natural-groups design with three conditions: morning types, intermediate types, and evening types (Vollmer & Randler, 2012). We would subsequently measure and compare the personality traits and values of the three groups. As another example, if we hypothesize that religiousness can have a positive influence on people's health, then we would measure people's religiousness (based on a particular operational definition) and use this to create a natural-groups design in which the different conditions reflect different degrees of religiousness (Lutgendorf et al., 2002). We would next measure and compare the health outcomes of the various groups. People's age, sex, and ethnicity are examples of subject variables frequently used to create natural groups (e.g., to study age, sex, and ethnic differences).

Is a Natural-Groups Design an Experiment? Two areas of disagreement arise concerning natural-groups designs. First, if a subject variable is the sole basis for creating different groups or conditions in a study, should we consider that subject variable to be an "independent variable"? Some researchers refer to such variables as *selected independent variables* or as *quasi-independent variables,* to distinguish them from *manipulated independent variables*. Others reserve the term *independent variable* only for variables that are manipulated. Personally, although I have no qualms about calling subject characteristics "independent variables" when discussing them conceptually, when discussing a single-factor natural-groups design operationally, I try to avoid doing so. But, many researchers do call them independent variables, and your course instructor may hold that view.

The second and more important issue is whether a single-factor natural-groups design constitutes an experiment. After all, it does seem as if the researcher is "tinkering around" with something by creating different groups. But in reality, a natural-groups design is a correlational study, not an experiment. As discussed in Chapter 5, in a correlational study the researcher measures variable X and variable Y, and then examines whether they are related statistically. Now reread the definition of a natural-groups design. They're the same, with one exception. To examine the relation between self-esteem and introversion–extraversion in a typical correlational study, we could give a self-esteem test to 200 students, and then correlate the actual self-esteem scores for all 200 students with their scores on a measure of introversion–extraversion. In a natural-groups design, we are not using each student's specific self-esteem score in the statistical analysis. Instead, we are using the score only to categorize the student into a low or high (or low, moderate, or high) self-esteem group. However, merely transforming the self-esteem data in this way does not transform the core methodology of a correlational study into an experiment. It doesn't change the fact that we measured people's naturally occurring self-esteem; we did not manipulate it.

It's important to understand that viewing a natural-groups design as a correlational study does not negate the importance of natural-groups designs. Correlational research has many important advantages and plays a vital role in advancing scientific knowledge. But, because a natural-groups design *is* correlational, its findings are open to the same two potential problems that limit the ability to draw clear causal conclusions from any correlational study: bidirectionality and the third-variable problem (see Chapter 5). For example, if we find that the low-self-esteem group, overall, is more introverted than the high-self-esteem group, the issue becomes: (1) Did lower self-esteem cause people to become more introverted, or did being introverted cause people to develop lower self-esteem (this is the bidirectionality problem); and (2) perhaps there is actually no causal relation between self-esteem and introversion; perhaps other personal or environmental factors either cause or are associated with lower self-esteem, and these other factors are the real causes of why people are more introverted (this is the third-variable problem). In sum, researchers need to be especially cautious when trying to draw causal conclusions from natural-groups designs.

RANDOM ASSIGNMENT VERSUS RANDOM SAMPLING

The ability to assign participants randomly to the different conditions of an experiment is a cornerstone of between-subjects experiments. Unfortunately, I've found that some students confuse random assignment with another important procedure, random sampling (discussed in detail in Chapter 7). Random assignment and random sampling do have some similarities: Both rely on the laws of probability and both are considered good scientific practice. But, they are used to achieve different goals. Here are some crucial differences:

- *Random sampling* is a procedure in which each member of a population has an equal probability of being selected into a sample chosen to participate in a study. In other words, random sampling is used to determine who will be asked to participate in a

study. Surveys are a common example: From the population of adults in a nation, a random sample would be selected to participate in the survey.

- *The goal of random sampling* is to select a sample of people whose characteristics are representative of the broader population from which those people have been drawn.
- *Random assignment* of participants, within the context of experiments, is used to determine the specific condition to which each participant will be exposed. Using random assignment, each participant has an equal probability of being assigned to any particular condition. Random assignment is not used to identify and select the sample of people who will be asked to participate in the experiment. Rather, once someone has agreed or been chosen to participate, random assignment determines the particular condition into which that person is placed.
- *The goal of random assignment* in experiments is to take whatever sample of people you happen to put together and place them in different conditions in an unbiased way. By doing this, at the start of the study researchers can assume that the groups of participants in the various conditions are equivalent, overall, to one another.

Here's another way to think about it. To select who will be in an experiment, random sampling is rarely used. Instead, researchers may place notices in newspapers or on websites to announce that they are seeking participants for a study. In many experiments, the participants are college undergraduates (e.g., students taking introductory psychology) who are recruited using sign-up sheets that are posted online or in a psychology department building. Suppose, therefore, that in an introductory psychology class of 150 students, 40 students sign up to participate in an experiment on decision making. In this case,

- the participants are college students, and they cannot automatically be considered a representative sample of the overall adult population;
- the students on this campus may not be representative of college students in general;
- the students enrolled in introductory psychology may not be representative of the overall student body on this particular campus; and
- the students who sign up for this particular experiment may not be representative of the entire class of students enrolled in introductory psychology.

If you have already read Chapter 7, *Survey Research,* you may recognize that the type of sampling used to recruit participants in this (and most) human psychology experiments is

Table 8.3 Differences Between Random Sampling and Random Assignment

	Random Sampling	Random Assignment (in experiments)
Description	Each member of a population has an equal probability of being selected into a sample chosen to participate in a study.	People who have agreed to participate in a study are assigned to the various conditions of the study on a random basis. Each participant has an equal probability of being assigned to any particular condition.
Example	From a population of 240 million adults in a nation, a random sample of 1,000 people is selected and asked to participate in a survey.	After a college student signs up for an experiment (e.g., to receive extra course credit or meet a course requirement), random assignment is used to determine whether that student will participate in an experimental or control condition.
Goal	To select a sample of people whose characteristics (e.g., age, ethnicity, gender, annual income) are representative of the broader population from which those people have been drawn.	To take the sample of people you happen to get and place them into the conditions of the experiment in an unbiased way. Thus, prior to exposure to the independent variable, we assume that the groups of participants in the various conditions are equivalent to one another overall.

called *convenience sampling*. Selection into the study is not done randomly; rather, experimenters in most cases take whomever they can get. As discussed in Chapter 7, convenience sampling is usually an adequate method of recruiting participants for experiments. It doesn't negate an experiment's ability to examine cause–effect relations. But for now, the main point is that whatever the characteristics of this particular sample of students may be, the researcher will use *random assignment* to determine who is placed in the various conditions of the experiment. In this way, we can consider the participants in the different conditions to be equivalent to one another, overall, at the start of the study. This enables us to examine whether our independent variable causes these initially equivalent groups to behave in different ways. **Table 8.3** summarizes major differences between random sampling and random assignment.

CONCEPT CHECK 8.3 BETWEEN-SUBJECTS DESIGNS

Decide whether each statement below is true or false. Answers appear on page 276.

1. To study some topics, a between-subjects design, rather than within-subjects design, must be used.
2. Overall, between-subjects designs require more participants than within-subjects designs.
3. In experiments, the purpose of random assignment is to select a sample of participants that is representative of the overall population.
4. In matched-groups designs, the matching variable and dependent variable must measure different characteristics.
5. Overall, between-subjects designs are less sensitive than within-subjects designs.

WITHIN-SUBJECTS DESIGNS

Learning Objectives

After studying this section, you should be able to:

- Describe some advantages and disadvantages of within-subjects designs.
- Explain the goals of counterbalancing.
- Describe several types of within-subjects designs.

How good are you at recognizing people's emotions from their facial expressions? Are different emotions equally easy (or difficult) to judge from facial cues? To elaborate on a possible emotion recognition experiment described earlier, let's now contrast two approaches to designing such an experiment. In both approaches, the independent variable is the type of emotion expressed and we'll create nine conditions: nine facial photographs of the same person (an actor) portraying anger, fear, joy, disgust, interest, surprise, contempt, shame, and sadness. Using a between-subjects design, there will be 180 participants, 20 per condition. After providing instructions to each participant, we'll show the participant one photograph, and in a few seconds that participant will make a judgment as to which emotion is being portrayed. Next we'll debrief the participant, who can then leave.

This between-subjects design is a legitimate approach to conducting the experiment, although in practice we would not want to limit ourselves to only nine photographs of the same actor. Maybe some of this particular actor's facial expressions will be uniquely easy or hard to judge. It would be better to have a set of nine photos for each of, say, 10 different actors, so that our data will be based on a broader sample of faces. But now we have 90 photographs, and if we ask 20 participants to judge each one, we'll need to have 1,800 participants in our study, each making a single judgment that takes only a few seconds.

You may be thinking that this is an inefficient way to conduct the experiment. Instead of asking 1,800 participants to each make a single judgment, why not ask 20 participants to judge all 90 photographs? This probably will take no more than half an hour per participant. Using this approach, we have now entered into the realm of *within-subjects designs* (also called *repeated-measures designs*), in which each participant engages in every condition of the experiment one or more times.

ADVANTAGES OF WITHIN-SUBJECTS DESIGNS

As the preceding numbers make clear, a key advantage of within-subjects designs is that they need fewer participants to obtain the same amount of data per condition than do between-subjects designs. This can be especially important in experiments where there are many conditions, the desired number of participants per condition is large, or the experimenter must spend extensive time giving instructions or leading each participant through a lengthy or difficult set of practice trials. The need for fewer participants also is important when they come from special populations (e.g., people with uncommon abilities or disorders) or are otherwise difficult to obtain.

Consider an experiment by Gabriel Kreiman and his colleagues that attempted to find the individual neurons in regions of the human brain that fire most frequently in response to specific types of complex visual stimuli (Kreiman, Koch, & Fried, 2000). The researchers identified these "category-specific" neurons by implanting microelectrodes inside 427 individual neurons within the brains of 11 patients with epilepsy. The microelectrodes would help doctors determine the neural source of the patients' seizures, and possibly lead to surgery to better control their epilepsy.

For the scientists, the patients' informed consent provided a remarkable research opportunity. Kreiman et al. (2000) recorded the firing rate of each neuron as the patient was presented with visual images that represented nine categories of stimuli, such as household objects, animals, cars, and abstract patterns. Multiple pictures were used for each category (there were several faces, cars, animals, etc.) and they were presented to each patient several times, resulting in up to 600 stimulus presentations for some patients. Each trial lasted 1 second. Given the large number of trials, the brevity of each trial, the relatively small number of participants available, and the delicate microelectrode implantation procedure performed on each patient, a within-subjects research design clearly was the only feasible way to conduct this particular experiment.

Another advantage of a within-subjects design is that, rather than recruiting a smaller number of participants to obtain the same amount of data per condition as a between-subjects design, we can instead recruit the same number of participants and thus collect more data per condition. In Figure 8.4 on page 253, the between-subjects and within-subjects designs for the longhand versus laptop note-taking experiment each have a total sample of 20 participants. Using a between-subjects design, we end up with data from only 10 students in each condition. Using a within-subjects design, we obtain twice the amount of data because each of the 20 students engages in both conditions. And, for our emotion recognition experiment, if 180 participants are available, then whereas a between-subjects design (with nine photographs, one per emotion) would only have 20 participants per condition, a within-subjects design would yield data from all 180 participants in each condition.

There is an important methodological and statistical advantage to having more participants per condition: Any findings we obtain are, in a sense, more reliable (i.e., less open to unusual chance fluctuations) because they are based on a larger amount of data. Other things being equal, if an independent variable truly has an influence on a dependent variable, an experiment will be more likely to detect that influence as the number of participants increases.

A third advantage is that within-subjects designs are the only approach that can be used to answer certain types of questions. For example, to examine whether specific neurons *change* their rate of firing in response to different stimuli, we have to measure those same neurons repeatedly as we expose a given participant to the different stimuli in each condition in our experiment. Similarly, in the field of *psychophysics,* within-subjects designs are required to investigate questions such as, "What is the smallest amount of change in the intensity of a sound that an individual can reliably perceive?" The only way to establish these individual perceptual thresholds—called *just-noticeable differences* or *difference thresholds*—is to expose each participant to the same sound played at varying decibels.

Finally, compared to between-subjects designs, within-subjects designs do a better job of creating equivalent groups at the outset of the experiment. Nothing can be more equivalent than having the same participants in all the conditions!

DISADVANTAGES OF WITHIN-SUBJECTS DESIGNS

Within-subjects designs run the risk that exposing participants to all the conditions will make them aware of the experiment's purpose or hypothesis. Moreover, either logically or practically, some research questions do not lend themselves to within-subjects experiments. Hess's (1959) imprinting experiment provides one example: As we discussed earlier, a duckling's first exposure to an adult model duck can only occur at a particular number of hours after birth. The experimenter can control when the initial exposure occurs, but the duckling can't possibly engage in more than one "first exposure" condition. Research on whether visual deprivation after birth impairs normal visual development provides another example. In such visual deprivation studies, newborn animals are raised from birth either in a dark or light environment for varying lengths of time, and later their performance in running a visually complex maze is compared (Crabtree & Riesen, 1979). An animal can grow up either exposed or never exposed to light, but cannot possibly grow up from birth in both types of conditions. Likewise, in an experiment comparing the effectiveness of behavioral, cognitive, and psychodynamic therapy, we would need to assign each participant to receive only one type of treatment.

When within-subjects designs are feasible, they still present a potentially huge problem: **order effects** (also called **sequence effects**), *which occur when participants' responses are affected by the order of conditions to which they are exposed*. In an experiment with conditions A, B, and C, if participants respond differently in condition A depending on whether it is the first, second, or third condition to which they are exposed, then this would be an example of an order effect.

Some order effects, called **progressive effects,** *reflect changes in participants' responses that result from their cumulative exposure to prior conditions*. For example, in an experiment where participants perform the same task under different conditions, they gain increasing task practice with exposure to each successive condition. In an experiment that examines the effects of cell phone conversations on driving performance, suppose that each participant drives the same simulated route once while using a handheld phone, once while using a hands-free phone, and once while not talking on a cell phone. As participants move from one condition to the next, they become more familiar with the driving route. This can create a **practice effect,** a *performance improvement due to greater experience with a task*. Conversely, participants might experience a **fatigue effect,** *a performance decline that results from becoming tired, inattentive, or less motivated to perform well with repeated exposure to a task*. Thus, if all participants in the cell phone experiment engage in the three conditions in the same order, and driving performance differs across those conditions, practice and fatigue effects could provide alternative explanations for any findings we obtain.

Other order effects, called **carryover effects,** *occur when participants' responses in one condition are uniquely influenced by the particular condition or conditions that preceded it*. In an experiment asking you to judge the sweetness of different drinks, a particular drink may seem more or less sweet to you depending on whether the drink that immediately preceded it was the sweetest or least sweet drink of all. Similarly, imagine a within-subjects laboratory experiment in which each participant, on three separate nights, is allowed 0, 4, or 8 hours of sleep and then performs a series of tasks at noon. Unless the researcher spaces each person's three trials far apart, carryover effects from sleep deprivation could occur. For example, if participation occurred on consecutive nights, how well people perform after 4 hours of sleep may differ depending on whether they received no sleep or 8 hours of sleep on the preceding night. As in judging the sweetness of drinks, the way participants react in a particular condition may be altered by the specific condition that comes just before it.

Within-subjects designs also may produce another type of order effect, **sensitization,** *in which exposure to multiple conditions increases participants' awareness of, or sensitivity to, the variable that is being experimentally manipulated*. For example, as Greenwald (1976) notes, in a within-subjects experiment designed to examine how different levels of room illumination influence workers' performance, awareness of changes in how brightly the room is lit may cause workers to become more sensitized to illumination as a factor affecting their performance. This may even lead participants to form hypotheses about how room illumination should affect their performance, and these expectations, in turn, might influence their performance. This type of order effect differs from a contrast effect, in which the perception of a room's brightness or dimness might be affected by the illumination level in the immediately preceding condition. Rather, sensitization focuses on participants' greater readiness to perceive any changes in the independent variable due to their multiple exposures to it.

In some experiments, such as those where the goal is to examine people's maximum sensitivity to detect changes in a stimulus (e.g., light or sound intensity), sensitization may facilitate the researcher's goal (Greenwald, 1976). In other experiments, such as those where researchers want to prevent participants from identifying the independent variable or minimize their awareness of it, sensitization effects from using a within-subjects design would be an undesirable factor.

Lastly, depending on the topic being studied, another potential disadvantage of within-subjects designs concerns the need to develop multiple, equivalent sets of materials to which participants are exposed as they progress through the different conditions of the experiment. The longhand versus laptop note-taking experiment portrayed in Figure 8.4 (p. 253) provides an example. As noted earlier, it makes little sense to have students watch the same lecture twice and take the identical knowledge test twice, once when taking longhand notes and once while taking laptop notes.

Therefore, we would want to develop two lectures that are equivalent on key characteristics, such as length, the overall number and difficulty of key concepts covered, the enthusiasm and clarity with which the lecturer delivers them, and the degree to which students find the two lectures to be engaging. Likewise, we would want to develop two knowledge tests—one for each lecture—that have the same number and type of items (e.g., factual, conceptual), and that are similar in overall difficulty. Accomplishing this would involve a process of developing an initial pair of lectures and knowledge tests, and assessing their equivalence by gathering data from a sample of students. We would then revise the materials and reassess their equivalence one or more times, based on additional feedback from new samples of students, until we have developed two lectures and knowledge tests that are similar in their key features. We would then initiate the actual note-taking experiment with a new sample of students.

COUNTERBALANCING AND TEMPORAL SPACING

In a within-subjects experiment, if all participants engage in the various conditions in the same order, then any differences in the dependent measure might be due to the influence of the independent variable, or instead, to the particular order in which the conditions appeared. Consider the longhand versus laptop note-taking experiment portrayed in Figure 8.4. Imagine that all students first take longhand notes during a lecture, and then use a laptop to take notes during a second lecture. Suppose that we find students' learning is better when they take longhand notes. Is this result due to the fact that longhand note-taking is the better approach for promoting learning, or might it be due to the fact that, after taking longhand notes for the first lecture, participants became bored or fatigued while taking laptop notes for the second lecture?

As a second example, consider the cell phone driving experiment in which every participant drives the same route first while conversing on a handheld phone, next while talking on a hands-free phone, and last while not using a cell phone. We find that driving performance is worst in the handheld condition, better in the hands-free condition, and best in the no-phone condition. Can we conclude that talking on a cell phone, especially a handheld phone, impairs performance? Clearly not: An obvious alternative explanation is that as participants progressed from the handheld to the hands-free to the no-phone conditions, they performed progressively better because they gained familiarity with the route.

To rule out such plausible alternative explanations, we must design our within-subjects experiment to avoid confounding factors that could arise by having all participants engage in the same order of conditions. To accomplish this we need to counterbalance the order of the conditions so that no one condition has an advantage or disadvantage relative to any other condition. Counterbalancing will not change the fact that order effects are likely to occur. Participants may still become fatigued, bored, more experienced, and so forth, as they move on to each subsequent condition. What counterbalancing does, however, is just what it says: It balances the sequence of the positions so there is less likelihood that order effects will work against or in favor of any particular condition.

On page 253, the right side of Figure 8.4 illustrates how we would counterbalance the sequence of conditions in the longhand versus laptop note-taking experiment. We would create two orders. For Order 1, the longhand note-taking condition comes first, followed by the laptop note-taking condition. For Order 2, the laptop note-taking condition comes first, followed by the longhand note-taking condition. With a sample of 20 students, 10 students would be randomly assigned to Order 1, and 10 students would be assigned to Order 2.

Moreover, even though we have created two equivalent lectures and knowledge tests for this experiment, we would want to apply a simple counterbalancing procedure to the order in which students view the lectures. Let's say that one lecture is about economics and the other is about geography. For the 10 students assigned to Order 1, five of them would begin the experiment by taking longhand notes during the economics lecture, and then switch to taking laptop notes for the geography lecture. The other five students assigned to Order 1 would first take longhand notes during the geography lecture, and then switch to laptop notes for the economics lecture. Similarly, for the 10 students assigned to Order 2, five would begin by taking laptop notes during the economics lecture, and switch to longhand notes for the geography lecture. The other five students assigned to Order 2 would begin by taking laptop notes for the geography lecture, and then switch to longhand notes for the economics lecture. Applying counterbalancing in this way would make it implausible for someone to argue that any results we obtained—such as superior conceptual learning when taking longhand notes—was due to a confound in the order of note-taking conditions or sequence in which the lectures were viewed.

In addition to counterbalancing, in some experiments it may be feasible to reduce certain order effects by allowing sufficient time to pass in between participants' exposure to the various conditions of the experiment. Fatigue effects are one example. If the dependent variable in an experiment is the quality of task performance, and participants will perform the same task in each of two experimental conditions, researchers may choose to set exposure to the conditions far enough apart to allow participants time to physically recover. The rest interval might vary from merely seconds to weeks, depending on the task. This procedure could also be used if the experimental manipulation itself would be expected to produce a fatigue effect. For instance, although many college students attend consecutive lectures without a break due to their course schedules, in our longhand versus laptop note-taking experiment we might consider providing participants with a 15- to 30-minute rest period after viewing the first of the two lectures.

As an example of longer temporal spacing, one research team used a within-subjects experimental design to examine whether short-term sleep deprivation caused people to act more impulsively (Cedernaes et al., 2014). Each participant stayed overnight in a sleep lab on two separate occasions. In one session they were permitted a normal night's sleep, and in the other they were deprived of all sleep. The order of the conditions was counterbalanced such that, as determined randomly, some participants were first exposed to the sleep deprivation condition and others were first exposed to the sleep condition. Additionally, for each participant the two sessions were spaced at least four weeks apart. Impulsivity was assessed in the morning of each session by measuring participants' performance (e.g., reaction time, number of errors) at a decision-making task. Sleep deprivation was found to increase impulsivity.

TYPES OF WITHIN-SUBJECT DESIGNS

Within-subjects designs fall into two general categories: those in which each participant is exposed to every condition in the experiment (1) only once and (2) more than once. To illustrate specific designs within each category, we'll examine how the same research question would be handled by each design.

Our featured example involves a long-running, multi-billion-dollar battle between corporate giants who are fighting for your mind, heart, and primarily your cash: the cola wars (**Figure 8.6**). And, the weapon of choice is the seemingly simple taste preference test. In the 1970s, PepsiCo, the makers of Pepsi-Cola, launched the famous Take the Pepsi Challenge campaign, still continuing today in modified form. For years, advertisements appeared in newspapers and magazines, based on taste-test experiments that PepsiCo conducted in different cities.

Suppose we want to design an unbiased experiment to see which of four non-diet cola drinks people prefer the most: Pepsi, Coke, Royal Crown Cola (RC), and Shasta. (Sorry, Dr. Pepper lovers: Your drink is not advertised as a "cola.") On each trial, participants will taste a drink and then rate it. We need to address one of the major possible confounding factors in such an experiment: the order of the drinks.

Daniel Acker/Bloomberg via Getty Images

Figure 8.6 The Pepsi–Coke cola wars. In taste tests and in advertisements, the battle between the makers of Pepsi and Coke for consumers' hearts and wallets has spanned decades.

Exposing Participants to Each Condition Once

We first examine three designs—*all possible orders*, the *Latin Square*, and *random selected orders*—in which every participant engages in all of the conditions one time.

All-Possible-Orders Design (Complete Counterbalancing). In an experiment with *n* conditions (where *n* stands for the number of conditions), there are *n*! ("*n* factorial") unique orders in which those conditions can be arranged. Thus, in our taste-test experiment, if we only had two drinks, Coke and Pepsi, there would be *2*! (i.e., $2 \times 1) = 2$ possible orders: Coke–Pepsi and Pepsi–Coke. If we had three drinks (say, adding RC Cola), there would be *3*! (i.e., $3 \times 2 \times 1) = 6$ possible orders:

Coke–Pepsi–RC	Pepsi–Coke–RC	RC–Coke–Pepsi
Coke–RC–Pepsi	Pepsi–RC–Coke	RC–Pepsi–Coke

And, if we have our full complement of four cola drinks (adding Shasta), we would have *4*! ($4 \times 3 \times 2 \times 1) = 24$ possible orders, as shown in **Table 8.4**.

With an **all-possible-orders design** (also called **complete counterbalancing**), *the conditions of an independent variable are arranged in every possible sequence, and an equal number of participants are assigned to each sequence*. Thus, in our taste-test experiment with four drinks, we would need a minimum of 24 participants and assign one participant to each of the 24 taste-testing sequences. We could use one of several techniques for deciding which order each participant follows, with simple random assignment being the most straightforward. Thus, the first participant in our taste test might receive the order RC–Coke–Pepsi–Shasta, the second participant might get Pepsi–Shasta–RC–Coke, and so on. We could also use this design with any multiple of 24 participants: 48, 72, 96, and so forth, depending on where we want to balance the "more data" versus "fewer participants" benefit of using a within-subjects design. If we had 48 participants, for example, then for participants 25 through 48, we would run through the sequence of all 24 possible orders again, using new random assignments to determine which participants get which orders.

The strong advantage of this design is that, because every possible order is used an equal number of times, every possible confounding effect involving the sequence of conditions is completely counterbalanced. This is why the term *complete counterbalancing* is used along with or instead of *all possible orders* to describe this design. Specifically, this design accomplishes three counterbalancing goals:

Goal 1 Every condition of the independent variable appears equally often in each position (e.g., in our example, each of the four drinks appears equally often in the 1st, 2nd, 3rd, and 4th position).

Goal 2 Every condition appears equally often before and after every other condition (e.g., Coke occurs immediately before Pepsi equally as often as Pepsi occurs immediately before Coke; this will be the same for any combination of two drinks you examine).

Goal 3 Every condition appears with equal frequency before and after every other condition, *within each pair of positions in the overall sequence* (e.g., looking at the 1st vs. 2nd position in the order, Coke occurs immediately before Pepsi as often as Pepsi

Table 8.4 All Possible Orders for a Single-Factor Design with Four Conditions

Coke–Pepsi–RC–Shasta	Pepsi–Coke–RC–Shasta	RC–Pepsi–Coke–Shasta	Shasta–Pepsi–RC–Coke
Coke–Pepsi–Shasta–RC	Pepsi–Coke–Shasta–RC	RC–Pepsi–Shasta–Coke	Shasta–Pepsi–Coke–RC
Coke–RC–Pepsi–Shasta	Pepsi–RC–Coke–Shasta	RC–Coke–Pepsi–Shasta	Shasta–RC–Pepsi–Coke
Coke–RC–Shasta–Pepsi	Pepsi–RC–Shasta–Coke	RC–Coke–Shasta–Pepsi	Shasta–RC–Coke–Pepsi
Coke–Shasta–RC–Pepsi	Pepsi–Shasta–RC–Coke	RC–Shasta–Coke–Pepsi	Shasta–Coke–RC–Pepsi
Coke–Shasta–Pepsi–RC	Pepsi–Shasta–Coke–RC	RC–Shasta–Pepsi–Coke	Shasta–Coke–Pepsi–RC

There are a total of 24 orders: Six orders begin with Coke, six with Pepsi, six with RC, and six with Shasta. Therefore, we would need a minimum of 24 participants—one per order—to use this complete design.

occurs immediately before Coke; it's the same for the 2nd vs. 3rd position, and 3rd vs. 4th position, and this balancing will be the same for any pair of drinks you examine).

The biggest disadvantage of this design is that, as the number of conditions increases, the number of participants needed rises rapidly to exceed what the experimenter is likely to obtain. Adding a fifth condition (e.g., a fifth cola drink) to our taste test increases the number of possible orders to 120. With six conditions, we would need 720 participants, and with seven, 5,040 participants! Thus, depending on the number of participants available to the experimenter, this design is most commonly feasible with only two, three, or four conditions; maybe five if access is available to a huge subject pool.

Latin Square Design. Before defining this design, it will be best to consider an example. Look at the matrix in **Table 8.5**. Above the top row, we have the four possible positions, also called *trials,* in the series of drinks: 1st, 2nd, 3rd, and 4th position.

Inside the matrix there are four rows, with each row representing a different order (i.e., sequence) of drinks. If you look across the four rows and then down the four columns (for positions 1 through 4), you will see that each drink appears only once in each row and in each column. Welcome to the Latin Square, a sort of two-dimensional Rubik's cube. Any Latin Square—and there are many types—can be defined as follows. In a single-factor experiment where the number of conditions of the independent variable equals *n,* a **Latin Square** *is an* n *(number of positions in a series)* × n *(number of orders) matrix in which each condition will appear only once in each column and each row.* For example, if we had eight drinks, this would yield an 8 (positions) × 8 (orders) matrix.

For any given number of conditions (say, four drinks), there are multiple ways to design a Latin Square. The design in Table 8.5 is sometimes called a Williams Square, and its beauty is its remarkable efficiency. For an independent variable with four conditions, rather than using the 24 orders that would be required with an all-possible-orders design, the *Williams Latin Square* uses only four orders to accomplish the two most important counterbalancing goals achieved by the all-possible-orders design:

Goal 1 Every condition of the independent variable appears equally often in each position (e.g., in our example, each of the four drinks appears equally often in the 1st, 2nd, 3rd, and 4th position).

Goal 2 Every condition appears equally often before and after every other condition (e.g., Coke occurs immediately before Pepsi as often as Pepsi occurs immediately before Coke; this will be the same for any combination of two drinks you examine).

The only goal that the Williams Latin Square design does *not* accomplish is having each condition appear before and after every other condition between each pair of positions in the overall sequence. So, for example, Pepsi comes immediately before Shasta once, and vice versa, but the Pepsi–Shasta order occurs in the position-1-to-position-2 sequence (see row 1), whereas the Shasta–Pepsi order comes in the position-3-to-position-4 sequence (see row 3).

Table 8.5 Latin Square Using a Williams Design

Participant	Trial 1	Trial 2	Trial 3	Trial 4
Bria	Pepsi	Shasta	Coke	RC
Jinsoo	Shasta	RC	Pepsi	Coke
Beatriz	RC	Coke	Shasta	Pepsi
Tamara	Coke	Pepsi	RC	Shasta

Still, this often minor type of order imbalance is a small price to pay for the efficiency of the Williams Latin Square design.

The fact that our four-drink Latin Square has only four orders, one per participant, does not mean that we are restricted to only four participants in our study. We can have any multiple of four participants: 8, 12, 16, 40, 80, or more participants if we wish. Ideally, we would not use this specific Latin Square over and over for each set of four participants. Rather, if possible, we would use a different Williams Latin Square for each set of four participants to increase the variety of specific orders that will occur, overall, throughout our experiment.

The chief limitation of the Williams Latin Square is that when an independent variable has an odd rather than even number of conditions, you cannot construct a single Latin Square that will achieve counterbalancing Goal 2, described earlier. With a 5×5 matrix, each of the five drinks will appear equally often in the 1st, 2nd, 3rd, 4th, and 5th positions; but it is impossible, say, for Coke to immediately come before Pepsi 2½ times, and for Pepsi to immediately precede Coke 2½ times. The solution is to construct a pair of 5×5 Latin Squares that together will achieve counterbalancing Goal 2: In one matrix, Pepsi–Coke will occur more often than Coke–Pepsi, but in the other square this will be reversed. The use of two Latin Squares will require us to double the number of participants, if each participant is to be exposed to every condition only once.

Most Latin Square designs don't require this special doubling-up of matrices to accommodate an odd number of conditions. **Table 8.6** shows a Latin Square created with a method called *random starting order with rotation*. In this method, we randomly select an order for the top row, and then rotate the position of each condition by one placement in each subsequent row. Thus, Coke moves from position 1 in row 1 to position 2 in row 2; Pepsi moves over to position 3; Shasta moves to position 4; and RC, which was last in row 1, swings over to the beginning of row 2. You can see that this approach continues as we move on to rows 3 and 4. With a fifth drink, we would simply have one more column (i.e., Trial 5) and one more row.

As with all Latin Squares, Goal 1 is achieved. Each drink appears only once in each position. But with a design based on a random starting order with rotation, counterbalancing Goal 2 and Goal 3 are not achieved. Coke, for example, immediately comes before Pepsi three times, but Pepsi never immediately precedes Coke. And without this, they can't possibly come immediately before and after each other within any pair of positions.

Random-Selected-Orders Design. Particularly when the number of conditions is large and thus the total number of possible orders greatly exceeds the number of participants available, another approach that some researchers use to counterbalance order effects is called a **random-selected-orders design:** *From the entire set of all possible orders, a subset of orders is randomly selected and each order is administered to one participant*. Realize that with this type of simple randomization, just as you are unlikely to get exactly 50 Heads in 100 flips of a coin, in our taste test it is unlikely that each of our four drinks will end up in the first position or any other position exactly 25% of the time, or will occur immediately before and after any other particular drink an equal number of times. Instead, the researcher is relying

Table 8.6 Latin Square Using a Random Starting Order with Rotation

Participant	Trial 1	Trial 2	Trial 3	Trial 4
Emaan	Coke	Pepsi	Shasta	RC
Brandon	RC	Coke	Pepsi	Shasta
Juan	Shasta	RC	Coke	Pepsi
David	Pepsi	Shasta	RC	Coke

on the laws of probability to create a set of orders that are unlikely to produce a statistically significant bias or advantage in favor of any particular condition relative to the others. Thus, by chance, Drink A may appear more often in position 1 than Drink C, but the magnitude of this difference is unlikely to be statistically significant. To be most effective, this counterbalancing approach should not be used when the number of participants is small; with only a few participants, there will not be enough random orders selected to let chance have a good opportunity to balance order effects.

Exposing Participants to Each Condition More Than Once

Before describing two designs of this type, perhaps you're wondering, "Why on earth would a researcher ever need or want to expose participants to each condition more than once?" There are at least three reasons why researchers would choose this approach:

- *For practical reasons*. In some cases, the experimenter may have access to only a small number of participants, too small to effectively use any of the counterbalancing designs we discussed earlier.
- *To examine the reliability (consistency) of participants' responses*. If the same participants taste the same set of four drinks again, would their ratings of the drinks be consistent with their earlier ratings? Would they again select the same drink as being the best? If we are interested in such questions, we will need to expose the same people to each condition multiple times.
- *To extend the generalizability of the results*. This involves representing each condition with multiple stimuli, rather than just one stimulus.

To illustrate this third reason, suppose we want to test whether people prefer, overall, the taste of cola drinks or that of clear "lemon-lime" soda drinks. We decide to use Pepsi and 7UP to represent each type of drink, and counterbalance the orders. If people prefer Pepsi to 7UP, have we really shown that they like cola drinks more than lemon-lime drinks? Perhaps if we had used Coke and Sprite, or Shasta and Sierra Mist, the results would have been different. To avoid this problem, we can choose several colas and several lemon-lime sodas. We expose each participant to the two conditions of our independent variable (cola vs. lemon-lime drink) several times, but each time the conditions involve different cola and lemon-lime drinks. Across the study, we counterbalance the order of drinks. Now, if people choose cola drinks more often, we are more confident in concluding that they generally prefer cola to lemon-lime drinks.

Block-Randomization Design. Imagine that the first participant, Teresa, shows up for our cola taste test. We administer each of the four colas to her one time, using a randomly selected order: RC–Shasta–Coke–Pepsi. This represents Block 1. Then we administer the four colas again, based on a fresh randomly selected order. So, for Block 2, suppose the order of drinks is Coke–Shasta–Pepsi–RC. Then (assuming that we have very nice, patient participants) we administer Block 3 to Teresa, once again randomly determining the order of drinks. We end with Block 4, which uses a freshly randomized order. When our second participant, Shih-Fen, arrives, she will also receive four blocks. And, for each of her blocks, the order of drinks will be randomized.

Earlier in the chapter, we discussed how block randomization is used in between-subjects designs to randomly assign different participants to the various conditions of an experiment (pp. 256–257). Here, as there, our entire set of conditions (in this case, the four types of cola drinks) is called a block, and the order of conditions in every block is randomly determined. However, block randomization is applied differently in within- versus

between-subjects designs. In a between-subjects design, each participant only engages in a total of one condition within one particular block. In contrast, in the within-subjects design we are now discussing, each participant not only performs all the conditions within a block, but also is exposed to multiple blocks. More specifically, in a **block-randomization design,** *every participant is exposed to multiple blocks of trials, with each block for each participant containing a newly randomized order of all the conditions.* A portion of a block-randomization design appears in **Table 8.7**, with the four drinks abbreviated as C (Coke), P (Pepsi), S (Shasta), and R (RC).

Using this design, if the number of blocks must be small—due, for example, to the amount of time or effort that each participant must expend for each trial of the task—then a greater number of participants will be needed in order to give randomization the opportunity to counterbalance order effects over the entire experiment. But even with a small number of participants, if there are many blocks per participant, then block randomization can successfully counterbalance order effects across the entire experiment. With our taste-test experiment, we might be pushing participants' patience by having more than three or at most four blocks. However, in some block-randomized experiments, participants may engage in 15 or 20, or possibly more, blocks of trials.

Reverse-Counterbalancing Design. With a **reverse-counterbalancing design** (also called an **ABBA-counterbalancing design**), *each participant receives a random order of all the conditions, and then receives them again in the reverse order.* Even though an independent variable may have more than two conditions (condition A and condition B), the term *ABBA* is used to signify the mirror-image nature of this design. **Table 8.8** gives an example of an ABBA reversal design. Thus in our taste test, for our first participant, Zach, we randomly determine the order of the four colas (say, Pepsi–Shasta–Coke–RC) and then reverse the sequence so that Zach will receive the following overall sequence: Pepsi–Shasta–Coke–RC–RC–Coke–Shasta–Pepsi. We could stop there or have Zach go through another set, in which we start with a fresh random order and then reverse it. We could repeat this cycle more times if we

Table 8.7 Portion of a Within-Subjects, Block-Randomization Design

Participant	Block 1	Block 2	Block 3	Block 4	Block 5
Teresa	R S C P	C S P R	C P R S	S P R C	R P C S
Shih-Fen	C P R S	P R C S	R S C P	S R P C	P R S C
Rachel	S R P C	R S P C	R P S C	P R C S	R S P C
Averi	S C P R	C R P S	P C S R	R P C S	S C P R

Note: C = Coke; P = Pepsi; R = Royal Crown; S = Shasta.

Table 8.8 Portion of an ABBA Reversal Design

	Trial															
Participant	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Zach	P	S	C	R	R	C	S	P	S	P	R	C	C	R	P	S
Miguel	R	S	P	C	C	P	S	R	P	R	C	S	S	C	R	P
Doug	C	P	R	S	S	R	P	C	C	S	R	P	P	R	S	C
Jon	R	C	S	P	P	S	C	R	R	S	C	P	P	C	S	R

Note: C = Coke; P = Pepsi; R = Royal Crown; S = Shasta.

wish, each time beginning with a new random order. For our next participant, Miguel, we use the same approach, each time starting with a new random order.

The logic behind reverse counterbalancing is that on average, for any ABBA sequence, each condition is guaranteed to end up with the same average position. So in our example, for Zach, in Trials 1 through 8, Pepsi is 1st and 8th, Shasta is 2nd and 7th, Coke is 3rd and 6th, and RC is 4th and 5th; thus, each drink has an average position of 4.5. This will be the same for every set of eight trials, for every participant. This is a rather different approach to counterbalancing, as opposed to trying to ensure that each drink appears equally often in each position. In Table 8.8, you can see that in three of the eight sets of trials shown, the first drink is RC Cola. You can see that Shasta comes in second in four of the eight sets. Now, if our experiment has considerably more participants or sets of trials per participant, then these types of imbalances become less likely to occur as randomization has a greater chance to operate. But as the number of sets per participant increases, this design—unlike pure block randomization—may enable participants to detect the ABBA pattern and thus start to anticipate the order of the conditions (at least in the last half of each series), or even to form hypotheses about the purpose of the experiment. Depending on the behavior being studied, such *anticipation effects* might alter the speed or nature of participants' responses and consequently begin to confound the results.

The notion that counterbalancing will be effective if the average position of the different conditions is kept equal is based on an assumption that the order effects which accrue with each trial add up in linear fashion. For example, in our taste test, with every drink tasted participants may get a touch more "fatigued" by the generally sweet taste of colas. Thus, a drink's advantage of being tasted first is balanced out by the disadvantage of being tasted last, to a similar degree that this balances out for drinks in positions 2-7, 3-6, and 4-5.

But suppose, analogous to our general discussion of nonlinear effects earlier in the chapter, that in a taste test the following happens for most people, as illustrated by Zach. For the first six tastes that Zach performs—Pepsi, Shasta, Coke, RC, RC, and Coke, everything is fine. Then, with just two drinks to go in the first set, Zach "hits the wall" and suddenly begins to find the sweet tastes of Shasta and Pepsi very bothersome, or sickeningly sweet. In this case, coming in position 7 and position 8 in the order takes on a substantially greater disadvantage that is not balanced out by whatever benefit might have accrued from being position 2 or position 1. In other words, if order effects build up in a nonlinear fashion, then the central assumption behind reverse counterbalancing doesn't work. This problem is called *nonlinear order effects*. There are several strategies for trying to minimize the confounding effect of nonlinear order effects in an ABBA design, but the researchers also might consider whether some other counterbalancing approach, such as the block-randomization design described above, would be more appropriate.

CONCEPT CHECK 8.4 WITHIN-SUBJECTS DESIGNS

Fill in each of the blanks below. Answers appear on page 276.

1. Collectively, within-subjects designs are also called _____ designs.
2. In within-subjects designs, two common types of progressive order effects are _____ effects and _____ effects, and experimenters use _____ to minimize order effects as possible confounding variables.
3. It is not feasible to use an all-possible-orders design when an experiment has many _____.
4. In a within-subjects _____ design and _____ design, each participant engages in each condition more than once.

EXAMINING THE RESULTS: GENERAL CONCEPTS

Learning Objectives

After studying this section, you should be able to:

- Describe how descriptive statistics and inferential statistics help researchers examine the results of their studies.
- Discuss a typical approach used in single-factor designs to examine whether the findings are statistically significant.

To conclude this chapter, we'll briefly discuss the general approach that researchers use to examine the results from single-factor experiments. We'll focus on concepts and not on statistical formulas or computations.

Researchers use *descriptive statistics* to summarize their data. Reporting the mean scores of participants in the different conditions (as in Figures 8.1 and 8.7), and reporting the percentage of trials on which a particular response occurred (as in Figure 8.2), are examples of how researchers use descriptive statistics. Researchers then utilize *inferential statistical tests* to help them determine whether their findings are statistically significant, that is, unlikely to be due simply to chance. (See Statistics Module 1 for a further overview of descriptive and inferential statistics, and Module 9 for a discussion of statistical significance.)

In most experiments, researchers perform statistical tests designed for dependent variables that have been measured on an interval or ratio scale (as opposed to a nominal or ordinal scale). We'll focus on those tests here. The two most common statistical tests used to analyze interval- and ratio-scale data from experiments are the *t test* and the *analysis of variance*. A *t test* helps researchers determine whether the difference between the mean scores of two conditions is statistically significant. *Analysis of variance* (ANOVA) helps researchers determine whether the overall pattern of differences among the mean scores of the conditions is statistically significant. (See Statistics Modules 12 and 16 for more detail on the *t* test and ANOVA.)

When a single-factor experiment has only two conditions, either a *t* test or an ANOVA can be used. Moreover, when an experiment has only two conditions, there is only one step to perform: You directly compare the two conditions. The analysis will tell you the probability that the difference between the mean scores of the two conditions is due to chance. Traditionally, if less than a 5% probability exists that chance factors could be solely responsible for the results, then the finding is considered statistically significant.

When a single-factor experiment has three or more conditions, then the analysis may involve several steps. Typically, the first step is to perform an ANOVA to determine whether the overall pattern of findings is statistically significant. For example, **Figure 8.7** shows the

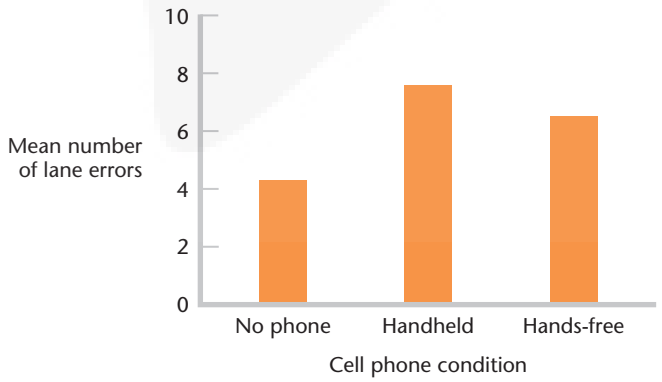


Figure 8.7 Hypothetical results from an experiment on cell phone use while driving.

results of a hypothetical experiment in which participants performed a driving simulator task while (1) not talking on a cell phone, (2) talking on a handheld cell phone, and (3) talking on a hands-free cell phone. The dependent variable is the number of lane errors: the number of times participants drove onto or across the white lane markers on either side of their lane. The ANOVA would determine whether, overall, the number of lane errors differs significantly depending on the phone condition.

Typically, if the ANOVA does not reveal a statistically significant overall pattern of findings, then the analysis stops. If the ANOVA is statistically significant, then researchers often proceed with tests that compare the means of specific conditions. Because these tests are performed after looking at the results and determining whether the overall pattern is significant, they are called *post-hoc tests* or *post-hoc comparisons*.

For example, if an ANOVA reveals that the overall pattern of results in Figure 8.7 is statistically significant, we can use a separate post-hoc comparison to determine whether there is a significant difference in the mean scores between (1) the no-phone versus handheld condition, (2) the no-phone versus hands-free condition, and (3) the handheld versus hands-free condition. Thus, our statistical analysis of a multilevel single-factor experiment has proceeded from an initial general analysis (ANOVA) to more specific comparisons between specific conditions (post-hoc tests). This is not the only data analysis approach that researchers can use, but it's a common one.

✓ CONCEPT CHECK 8.5 EXAMINING THE RESULTS: GENERAL CONCEPTS

Decide whether each statement below is true or false. Answers appear on page 276.

1. If an experimental finding is statistically significant, this means that it is unlikely the finding is due solely to chance.
2. To determine whether a finding is statistically significant, a researcher would use inferential statistics.
3. Usually, if an experiment has three or more conditions, the initial data analysis will examine whether the overall pattern of findings is significant, rather than examine differences in every possible pair of conditions.

CHAPTER SUMMARY

- Experiments provide the ability to manipulate an independent variable, to control how and when the dependent variable is measured, and to regulate extraneous factors. This control offers the best opportunity to examine how independent variables causally influence dependent variables.
- In between-subjects designs, potential confounding variables that can arise from differences in participants' characteristics are addressed by using random assignment to create equivalent groups at the start of the experiment. In within-subjects designs, such potential confounding variables are addressed by having the same people participate in every condition of the experiment and counterbalancing the order of conditions.
- In a single-factor design, the researcher manipulates one independent variable. An independent variable can have two levels, but more levels are needed to determine whether the relation between an independent and dependent variable is linear or nonlinear.
- An experimental group receives a treatment or an "active" level of the independent variable. A control group receives no treatment or a baseline level of an independent variable. In some experiments, the concept of a control group does not apply.
- Between-subjects designs have several advantages. There is no possibility of order effects and they are less likely than within-subjects designs to tip off participants about the hypothesis or underlying purpose of the experiment. However, they are less

effective than within-subjects designs in creating equivalent groups and require more participants.

- In an independent-groups design, participants are randomly assigned to the various conditions. Block randomization ensures that the number of times each condition is conducted stays in lockstep with every other condition.
- Using a matched-groups design, sets of participants are matched on one or more variables before being randomly assigned to the various conditions of the experiment. The matching variable may be the same variable that constitutes the dependent measure, or it may be another variable that we are concerned about as a possible confounding factor.
- In a natural-groups design, the "conditions" are created by sorting people into different categories based on a naturally occurring subject variable. A natural-groups design is fundamentally a correlational study, not an experiment.
- Random assignment differs from random sampling. Random sampling is used to select the sample of individuals who will be asked to participate in a particular study. Random assignment is used in experiments to assign participants to one condition or another.
- Compared to between-subjects designs, within-subjects (repeated-measures) designs need fewer participants to obtain the same amount of data and do a better job of creating equivalent groups. They introduce the major potential confounding factor of order effects, which is addressed by counterbalancing the order of conditions.
- The all-possible-orders design, Williams Latin Square design, and random-selected-orders design are counterbalancing approaches that expose each participant to each condition just once. Block randomization and ABBA designs are counterbalancing approaches that expose each participant to each condition more than once.
- The *t* test and analysis of variance (ANOVA) are the most common statistical tests used to analyze data from single-factor designs. Either test can be used if the design only has two conditions. For multilevel designs, an ANOVA determines whether the overall pattern of differences among the mean scores is statistically significant. If it is significant, then post-hoc tests can be used to compare the means of specific conditions.

KEY TERMS

all-possible-orders design (complete counterbalancing) (p. 267)
between-subjects design (p. 247)
block randomization (p. 257)
block-randomization design (p. 271)
carryover effects (p. 264)
confounding variable (p. 245)
control condition (control group) (p. 251)
counterbalancing (p. 247)
dependent variable (p. 243)
experiment (p. 243)
experimental control (p. 244)
experimental condition (experimental group) (p. 251)
extraneous variable (p. 245)
fatigue effect (p. 263)
independent-groups design (random-groups design) (p. 255)
independent variable (p. 243)
Latin Square (p. 268)
matched-groups design (p. 257)
matching variable (p. 257)
natural-groups design (p. 258)
order effects (sequence effects) (p. 263)
practice effect (p. 263)
progressive effects (p. 263)
random assignment (p. 247)
random-selected-orders design (p. 269)
reverse-counterbalancing design (ABBA-counterbalancing design) (p. 271)
sensitization (p. 264)
single-factor design (p. 248)
subject variable (p. 258)
within-subjects design (p. 247)

ASSESS YOUR KNOWLEDGE

1. Describe the three key components of experimental control.
2. How do the components of experimental control enable experiments to satisfy the three key criteria for inferring cause and effect.
3. Identify and illustrate two general sources of potential confounding variables in experiments.
4. One way to create an independent variable is to manipulate some aspect of the physical

environment. Describe at least three other ways to create independent variables.

5. Provide some examples of quantitative and qualitative independent variables.
6. When deciding how many levels of an independent variable to create, what factors do researchers consider? How does the desire to examine nonlinear effects impact this decision?
7. What is a control group? Do all experiments have a control group? Explain.
8. How do between-subjects and within-subjects designs differ? What are some advantages of between-subjects designs?
9. How is block randomization used in an independent-groups design?
10. Explain the procedure for creating a matched-groups design, including two general ways to select a matching variable. What are some pros and cons of using matching versus an independent-groups design?
11. Explain some key differences between random assignment and random sampling.
12. What is a natural-groups design? Explain why a natural-groups design is or is not an experiment.
13. Describe some advantages and disadvantages of within-subjects designs.
14. Identify some specific types of order effects. What general approach is used in within-subjects designs to control for order effects?
15. Identify the names of five specific types of within-subjects designs.
16. Describe the procedure used in the all-possible-orders, Latin Square, and random-selected-orders designs.
17. Describe the procedure used in the block randomization and reverse counterbalancing within-subjects designs. As a group, what is the one key way in which these two designs differ from the all-possible-orders, Latin Square, and random-selected-orders designs?
18. Contrast the pros and cons of using within-subjects designs that expose each participant to each condition only once, versus more than once.
19. Discuss two general steps in analyzing the results of a single-factor design that has three or more conditions.

CONCEPT CHECKS: ANSWERS

8.1 The Logic of Experimentation

1. c **2.** d **3.** b **4.** f **5.** e **6.** a

8.2 Manipulating Independent Variables

1. false **2.** false **3.** true **4.** true **5.** true

8.3 Between-Subjects Designs

1. true **2.** true **3.** false **4.** false **5.** true

8.4 Within-Subjects Designs

1. repeated measures **2.** pratice; fatigue; counterbalancing **3.** conditions **4.** block randomization; reverse counterbalancing

8.5 Examining the Results: General Concepts

1. true **2.** true **3.** true

THINKING CRITICALLY AND APPLYING YOUR KNOWLEDGE

EXERCISE 1 Analyze the Experiment

The designs for two experiments are specified below. For each experiment, identify:

(a) the independent and dependent variable;

(b) whether the experiment represents a between- or within-subjects design;

(c) the specific type of between- or within-subjects design.

Bowl Me Over: The Original

Dr. Tindale recruits 30 women to participate in a laboratory experiment. None has ever bowled or played video bowling games. Using a Nintendo Wii U system, each woman bowls the virtual ball five times to get an idea of what the task involves. Then each woman bowls one complete game under each of two conditions: (1) alone (no audience is present) or (2) an audience of four men silently watches. Half the women bowl alone

first, and then bowl with an audience watching. The other half bowls their first game with the audience watching, and their second game with no audience present. Dr. Tindale records how well participants perform (i.e., records their bowling scores) in each condition.

Bowl Me Over: The Sequel

Dr. Tindale conducts a second experiment, with the following changes. There are 60 new participants, each of whom bowls only one game either alone or in front of a four-person or an eight-person audience. The first three participants are randomly assigned, one to each of the three conditions. Then the next three participants are randomly assigned, such that each condition now has a total of two participants. This continues until there are 20 participants in each condition. Dr. Tindale records how well participants perform (i.e., records their bowling scores).

EXERCISE 2 To Be, or Not To Be—Manipulated

Three studies are described below. Answer the following questions for each one:

(a) Conceptually, what are the independent and dependent variables? How are they operationally defined?

(b) Is the independent variable manipulated or selected?

(c) Is the proposed study an experiment? Why or why not?

Place Your Bets

Dr. Hendricks uses a psychological test to identify 30 extraverted and 30 introverted college students. In a lab, students each receive $50 ($5 for each of 10 trials) to use in a gambling game, in which they choose the size of their bet on each trial. Students keep whatever money remains at the end of the experiment. To measure their degree of risk-taking, Dr. Hendricks records the size of each bet. She finds that, compared to introverts, extraverts make larger bets.

I Want to Eat

Dr. Denorfia conducts a hunger study. Participants (one per day) eat their normal breakfast at home between 8:30 a.m. and 9:00 a.m., and don't eat again until they come to the lab at 5 p.m. At the lab, the participants rate how hungry they feel and, based on this, are classified into a low, moderate, or high hunger condition. They then are exposed to a buffet that offers 15 types of foods and are permitted to eat for 10 minutes. Dr. Denorfia records the total amount of food (i.e., measured by the weight of food) and number of different foods that each participant eats. He finds that as hunger increases, people eat more food but not a greater variety of food.

Seriously, I Want to Eat

Dr. Watanabe conducts a study on hunger. Each participant comes to the lab at 8:30 a.m. and is fed the same breakfast, ending at 9 a.m. Participants are randomly assigned to remain in the laboratory for either 3, 6, or 9 hours, during which time they are not permitted food. Based on this deprivation period, they are considered to represent low, moderate, and high hunger groups, respectively. At the end of the deprivation period, each participant is exposed to a buffet that consists of 15 different types of foods and is allowed to eat for 10 minutes. Dr. Watanabe records the amount of each food that each participant eats. She finds that as their level of hunger increases, people eat more food but not a greater variety of food.

EXERCISE 3 The Face of Emotion

Dr. Goodman studies people's brain activity in response to seeing facial expressions that signal different emotions. Each participant is shown different photographs of a person's face, with each photo portraying a different emotion:

Joy (J) Anger (A) Fear (F) Sadness (S)

In the first and fourth experiments, three emotions are portrayed. In the other experiments, four emotions are portrayed. For each experiment, identify the specific type of within-subjects design that was used.

Experiment 1

	Trial		
	1	2	3
Participant 1	J	A	S
Participant 2	J	S	A
Participant 3	A	S	J
Participant 4	A	J	S
Participant 5	S	J	A
Participant 6	S	A	J

Experiment 2

	Trial			
	1	2	3	4
Participant 1	J	S	A	F
Participant 2	S	F	J	A
Participant 3	F	A	S	J
Participant 4	A	J	F	S

Experiment 3

	Trial			
	1	2	3	4
Participant 1	A	J	F	S
Participant 2	S	A	J	F
Participant 3	F	S	A	J
Participant 4	J	F	S	A

Experiment 4

	Trial					
	1	2	3	4	5	6
Participant 1	J	A	S	S	A	J
Participant 2	J	S	A	A	S	J
Participant 3	A	S	J	J	S	A
Participant 4	S	A	J	J	A	S
Participant 5	S	A	J	J	A	S
Participant 6	J	S	A	A	S	J