Experimental Design I: Single-Factor Designs

7

PREVIEW & CHAPTER OBJECTIVES

Chapters 5 and 6 have set the stage for this and the following chapter. In Chapter 5, we introduced you to the experimental method; distinguished between independent, extraneous, and dependent variables; considered the problem of confounding; and discussed several factors relating to the validity of psychology experiments. Chapter 6 compared between-subjects and within-subjects designs, described the basic techniques of control associated with each (e.g., random assignment, counterbalancing), and dealt with the problems of experimenter and subject bias in psychological research. With the stage now ready, this and the next chapter can be considered a playbill—a listing and description of the research designs that constitute experimental research in psychology. This chapter considers designs that feature single independent variables with two or more levels. Adding independent variables creates factorial designs, the main topic of Chapter 8. When you finish this chapter, you should be able to:

- Identify and understand the defining features of the four varieties of single-factor designs: independent groups, matched groups, ex post facto, and repeated measures.
- Describe two reasons for using more than two levels of an independent variable.
- Decide when to use a bar graph to present data and when to use a line graph.
- Describe the goals of the Ebbinghaus memory research, his methodology, and the results he obtained.
- Understand the logic behind the use of three types of control groups: placebo, wait list, and yoked.
- Understand the ethical issues involved when using certain types of control groups.
- Know when to use an independent samples *t*-test and when to use a dependent samples *t*-test, when doing an inferential analysis of a single-factor, two-level design.
- Understand why a one-way ANOVA, rather than multiple *t*-tests, is the appropriate analysis when examining data from single-factor, multilevel studies.
- Understand why post hoc statistical analyses typically accompany one-way ANOVAs for single-factor, multilevel studies.

In Chapter 3's discussion of scientific creativity we used the origins of maze-learning research as an example. Willard Small's research, using a modified version of the Hampton Court maze, was the first of a flood of studies on maze learning that appeared in the first two decades of the 20th century. Most of the early research aimed to determine which

of the rat's senses was critical to the learning process. You might recall from Box 2.3 of Chapter 2 that John Watson ran into trouble with "antivivisectionists" for doing a series of studies in which he surgically eliminated one sense after another and discovered that maze learning was not hampered even if rats were deprived of most of their senses. He concluded that rats rely on their muscle (kinesthetic) sense to learn and recall the maze. In effect, he argued that the rat learns to take so many steps, and turn right, and so on.

To test his kinesthesis idea directly, he completed a simple yet elegant study with his University of Chicago colleague Harvey Carr. After one group of rats learned a complicated maze, Carr and Watson (1908) removed a middle section of the maze structure, thereby making certain portions of the maze shorter than before, while maintaining the same overall maze design. They predicted that rats trained on the longer maze might literally run into the walls when the maze was shortened. Sure enough, in a description of one of the rats, the researchers noted that it "ran into [the wall] with all her strength. Was badly staggered and did not recover normal conduct until she had gone [another] 9 feet" (p. 39). A second group of rats was trained on the shorter maze and then tested on the longer one. These rats behaved similarly, often turning too soon and running into the sidewall of an alley, apparently expecting to find a turn there. Long after he left academia, John Watson remembered this study as one of his most important. Subsequent research on maze learning questioned the kinesthesis conclusion, but the important point here is that good research does not require immensely complex research designs. In some cases, comparing two groups will do just fine.

Single Factor—Two Levels

As you can see from the decision tree in Figure 7.1, four basic research designs can be called **single-factor designs**, with the term *factor* meaning "independent variable" here. Thus, single-factor designs have one independent variable. We will start with the simplest ones, those with two levels of the independent variable. The four designs result from decisions about the independent variable under investigation. First, the independent variable can be tested either between- or within-subjects. If it is tested between-subjects, it could be either a manipulated or a subject variable. If the independent variable is manipulated, the design will be called either an **independent groups design**, if simple random assignment is used to create equivalent groups, or a **matched groups design**, if a matching procedure followed by random assignment is used. As you recall from Chapter 6, decisions about whether to use random assignment or matching have to do with sample size and the need to be wary about extraneous variables that are highly correlated with the dependent variable.

If a subject variable is being investigated, the groups are composed of different types of individuals (e.g., male or female, introverted or extroverted, liberal or conservative, living in Pittsburgh or living in Cleveland). This design is called an **ex post facto design** because the subjects in the study are placed into the groups "after the fact" of their already existing subject characteristics. Researchers using ex post facto designs typically attempt to make the groups as similar as possible with reference to other variables. For instance, a study comparing males and females might select participants for each group that are about the same age and from the same socioeconomic class. Note that this type of matching, in which males are recruited so they are comparable in age and class to females, is a bit different from the kind of matching that occurs in the matched group design. In the latter, after matched pairs have been formed, they are randomly assigned to groups. In ex post facto designs, random assignment is not possible; subjects are already in one group or another by virtue of the subject variable being investigated (e.g., gender). You will see these two forms of matching in Research Examples 12 and 13 below.

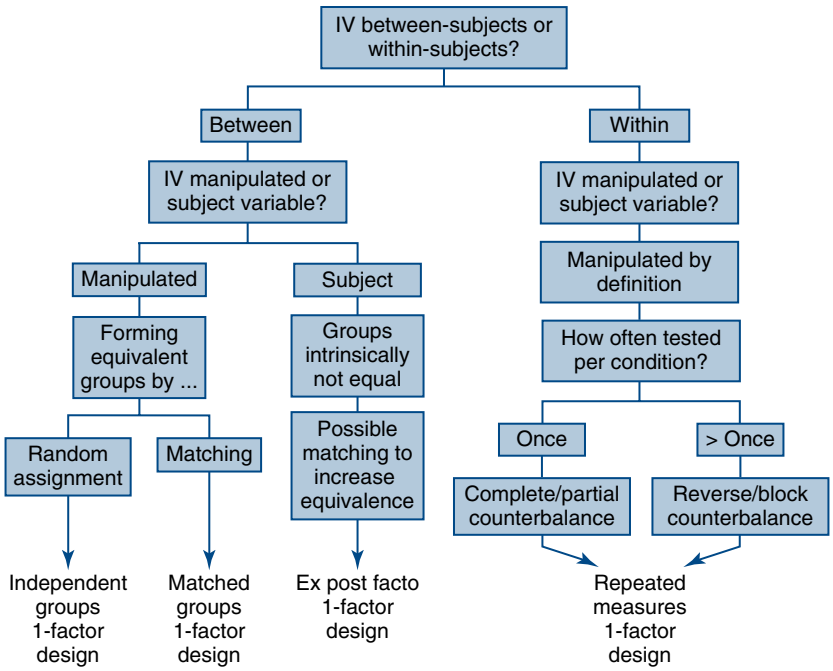

Figure 7.1
Decision tree—single-factor designs.

Table 7.1 Attributes of Four Single-Factor Designs

Types of Design	Minimum Levels of Independent Variable?	Independent Variable Between or Within?	Independent Variable Type?	Creating Equivalent Groups
Independent groups	2	between	manipulated	random assignment
Matched groups	2	between	manipulated	matching
Ex post facto	2	between	subject	matching may increase equivalence
Repeated measures	2	within	manipulated	n/a

The fourth type of single-factor design is a **repeated-measures design**, used when the independent variable is tested within-subjects—that is, each participant in the study experiences each level of the independent variable (i.e., is measured repeatedly). The major attributes of each of the four main types of designs are summarized in Table 7.1. Let's look at specific examples.

Between-Subjects, Single-Factor Designs

Single-factor studies using only two levels are not as common as you might think. Most researchers prefer to use more complex designs, which often produce more elaborate and more intriguing outcomes. Also, journal editors are often unimpressed with single-factor, two-level designs.

Nonetheless, there is a certain beauty in simplicity, and nothing could be simpler than a study comparing just two conditions. The following Research Examples illustrate three such experiments, one comparing independent groups, a second comparing matched groups, and a third comparing groups in an ex post facto design.

Research Example 11—Two-Level Independent Groups Design

An example of an independent groups design using a single factor with two levels is the first of an interesting set of studies by Mueller and Oppenheimer (2014) that investigated an important aspect of everyday student life—taking notes in class. In particular, Mueller and Oppenheimer wondered about the effectiveness of using laptops for note-taking purposes. We would guess that you like the idea of having information readily available on your devices, so laptop note-taking would seem to be a useful way to have important classroom information at your fingertips; the procedure might even enhance your learning of course material. Or so you might think—the evidence suggests that laptop note-taking is not very effective at all.

The defining feature of an independent groups design is random assignment of participants to groups, and that is what happened in the first of a series of three studies completed by Mueller and Oppenheimer (2014). Participants were assigned randomly to one of two groups. Those in the "laptop" group watched a 15-minute video lecture on one of several topics and were told to "use their normal classroom note-taking strategy" (p. 1160) while typing notes into their laptops. Those assigned to the "longhand" group saw the same lecture and were given the same instructions (i.e., normal note-taking strategy), but instead of using a laptop, they wrote their notes in longhand on note paper. After finishing the lecture, participants completed several distractor tasks lasting 30 minutes and then had their memory for the lecture tested. Questions (quoted from p. 1160) were both factual (e.g., "Approximately how many years ago did the Indus civilization exist?") and conceptual (e.g., "How do Japan and Sweden differ in their approaches to equality within their societies?").

On the factual questions, no differences occurred between the two groups. So, there was no advantage to taking notes on a laptop versus taking notes by hand when being tested on factual questions. On the conceptual questions however, there was a clear advantage for those taking notes by longhand over those typing notes into the laptop. This was true regardless of the topic of the lecture that participants listened to, and despite the fact that laptop note takers wrote significantly more words in their notes than did longhand note takers.

In Chapter 3 you learned that research outcomes often lead naturally to follow-up studies and that was exactly what happened with Mueller and Oppenheimer (2014). Using good "*what's next*" thinking and *conceptual replication* of their first experiment, they wondered what might account for the group differences. They had noticed in Study 1 that the laptop note takers seemed to transcribe the lectures verbatim, so they wondered if the advantages of longhand note taking might disappear if all subjects were explicitly instructed *not* to take down the lecture content in verbatim terms. What they found in Study 2 was that asking laptop subjects *not* to take notes verbatim did not improve memory performance and that these subjects still, despite the instructions, tended to write down content word for word. This outcome led to Study 3, in which they reasoned that perhaps the longhand note advantage would disappear if subjects were allowed to study their notes before taking the memory test (you can detect some *ecological validity* thinking here, creating a situation close to what students typically do—take notes, study notes, take a test). In this third study, longhand note takers still prevailed. Taken together, this set of experiments have clear implications for students in class—take notes by hand, and don't copy down what the professor is saying word for word. Translating lecture content into your own words produces a deeper level of information processing and improves memory for the lecture.

One final methodological point about this study is worth mentioning here. Especially on the conceptual questions, scoring was to some extent subjective, opening the door for some potential

bias. To control for this, the study's first author scored all the responses while not knowing which group was being scored; furthermore, a second scorer was used, and the results of the two scorers were compared. You might recognize this as a form of reliability (Chapter 4); in this case it is called **inter-rater reliability** and it was quite high in the study, giving the authors confidence that the test scoring was accurate.

Research Example 12— Two-Level Matched Groups Design

As you recall from Chapter 6, researchers often use a matching procedure, followed by random assignment, when (a) they have a small number of subjects, (b) they are concerned that some attribute of these subjects could affect the outcome, and (c) they have a good way of measuring that attribute. All three of those factors occurred in an interesting study on autism by Kroeger, Schultz, and Newsom (2007). They developed a video peer-modeling program designed to improve the social skills of young children (ages 4 to 6) with autism.

During 15 hour-long sessions over a 5-week period, some children with autism participated in a "Direct Teaching" group, spending part of their sessions seeing a brief video of other 4- to 6-year-old children modeling social skills that are typically absent in children with autism (e.g., taking turns with another child in a puzzle completion task). They then were given the opportunity to imitate those skills during a play period, with reinforcement for doing so. Children in a second ("Play Activities") group did not see the videos; they spent the time in play sessions that equaled the time the video-modeling group spent watching the video and then playing. Everything else was the same for both groups—for instance, children in the second group were also reinforced for displaying such social skills as taking turns in a puzzle task.

There were only 25 children in the study, and children with autism can display a wide range in their general level of functioning. Hence, a simple random assignment procedure might run the risk of producing two groups with different average levels of functioning. To avoid the problem, Kroeger et al. (2007) used a matching procedure followed by random assignment. Using a standard scale for measuring autism, the Gilliam Autism Rating Scale (a checklist completed by parents that lists a range of behaviors), Kroeger and her colleagues created pairs of children with matching levels of functioning (an "Autism Quotient"), and then randomly assigned one of each pair to each group.[1] The procedure worked; the average Autism Quotient scores for the two groups were essentially identical (92.15 and 92.58). Because it is necessary to have close supervision and small groups when working with children with autism, there were actually three Direct Teaching groups and three Play Activities groups, and a child-adult ratio of 2:1 was maintained in all groups.

In a study like this one, given the range of behaviors that can occur in a free play situation, it is important to have good *operational definitions* for the behaviors being measured. In this case, Kroeger et al. (2007) had a standardized measure available to them, the Social Interaction Observation Code. It measures "the prosocial behaviors of initiating a social interaction, responding to an initiation or invitation to socialize, and maintaining that social interaction" (p. 814). Even with such a tool, however, observations were made by fallible human observers, who did their scoring while watching video of the sessions. Three important control procedures were used to minimize human error and bias. First, observers were given extensive training. Second, researchers used the *double blind* procedure you learned about in Chapter 6; observers scoring a particular session did not know if the children were in the modeling group or not. Third, because any one observer might make errors, pairs of observers were used for each video segment, and the extent to which their coding results agreed was assessed. As in Mueller and Oppenheimer's

[1] Good for you if you wondered how the matching procedure worked with an odd number of subjects ($N = 25$). The senior author of the study (it was her doctoral dissertation) recalls that after 12 pairs had been formed, the parents of the remaining subject were given the choice of which group their child would join. Naturally, researchers don't like to turn away willing participants. Even with this slight deviation from a normal matching procedure, children in the two groups were equivalent on the matching variable, the Autism Quotient scores (K. A. Kroeger-Geoppinger, personal communication, November 8, 2011).

(2014) note-taking study, *inter-rater reliability* was assessed for observers who scored the video. After undergoing training, the observers in this autism study were quite good; they agreed 98.4% of the time. As for the results, the video modeling procedure worked. The children in the Direct Teaching group showed dramatic improvement in their social interaction skills, while those in the Play Activities group did not change.

Research Example 13— Two-Level Ex Post Facto Design

One type of research that calls for an ex post facto design examines the effects of brain damage that results from an accident. For obvious ethical reasons, an independent groups design with human subjects is out of the question (any volunteers for the experimental group, those randomly assigned to the group receiving the brain damage?). Although most research studies comparing those with traumatic brain injuries (TBI) look at cognitive factors (e.g., effects on memory or language), an interesting study by McDonald and Flanagan (2004) investigated the abilities of 34 subjects with TBI to process and understand social exchanges. As with many ex post facto studies, the researchers tried to select subjects so the two groups would be as similar as possible, except for the brain damage; in this case they selected control group subjects (without TBI) that were "matched on the basis of age, education, and gender" (p. 574). Note again that this is matching "after the fact" and different from the situation in Research Example 12, when matching was followed by random assignment. Matching followed by random assignment creates equivalent groups; matching in an ex post facto design makes the groups more similar to each other, but we cannot say equivalent groups are the result because random assignment is not possible with the subject variables that define the ex post facto design.

In the McDonald and Flanagan (2004) study, both groups viewed brief videos from The Awareness of Social Inference Test (TASIT). The videos portrayed people having various kinds of social interactions and displaying a range of emotions. For instance, one TASIT video includes an exchange in which one person is displaying sarcasm, while another video has one person lying to another. McDonald and Flanagan were interested in determining if those with TBI were impaired in their ability to (a) accurately detect the basic emotions being felt by those in the videos (e.g., the anger of someone who was being sarcastic), (b) distinguish sincere from sarcastic comments, and (c) distinguish "diplomatic" lies (told to avoid hurting someone's feelings) from lies told in a sarcastic fashion. The results? Compared to controls, those with TBI were significantly impaired in their abilities to recognize emotions and to recognize a lack of sincerity. For example, because they had problems detecting anger, they found it hard to distinguish sarcasm from sincerity.

In studies like this, one methodological concern is *external validity* (Chapter 5). To what extent did the 34 experimental subjects in the McDonald and Flanagan (2004) study represent a typical TBI patient? The researchers were aware of the issue and took pains to select participants who would, as a group, reflect the usual attributes of TBI patients. For example, they compared the number of days of posttraumatic amnesia (76) for their TBI subjects with the number of amnesia days reported "in a consecutive series of 100 people with TBI who were discharged from a comparable brain-injury unit in an independent study" (p. 573), and found no significant difference. From this and other factors, McDonald and Flanagan concluded their "group was representative of the severity of injury typically seen in this population" (p. 573).

Within-Subjects, Single-Factor Designs

As you already know, any within-subjects design (a) requires fewer participants, (b) is more sensitive to small differences between means, and (c) typically uses counterbalancing to control for order effects. A within-subjects design with a single independent variable and two levels will counterbalance in one of two ways. If subjects participate in each condition just once, complete counterbalancing will be used. Half of the participants will experience condition A and then B,

and the rest will get B and then A. If participants are tested more than once per condition, reverse counterbalancing (ABBA) could be used. This route was taken by J. Ridley Stroop in the first two of three studies he reported in 1935. This study is high on anyone's "Top 10 Classic Studies" list. For a close look at it, read Box 7.1 before continuing.

BOX 7.1 CLASSIC STUDIES—Psychology's Most Widely Replicated Finding?

Reverse counterbalancing was the strategy used in a study first published in 1935 by J. Ridley Stroop. The study is so well known that the phenomenon it demonstrated is now called the "Stroop effect." In an article accompanying a 1992 reprinting of the original paper, Colin MacLeod called the Stroop effect the "gold standard" of measures of attention, and opened his essay by writing that

> it would be virtually impossible to find anyone in cognitive psychology who does not have at least a passing acquaintance with the Stroop effect. Indeed, this generalization could probably be extended to all those who have taken a standard introductory course, where the Stroop task is an almost inevitable demonstration. (MacLeod, 1992, p. 12)

MacLeod went on to state that the Stroop effect is one of psychology's most widely replicated and most frequently cited findings—a PsycINFO search for "Stroop effect OR Stroop test OR Stroop interference" yields more than 4,000 hits. So what did J. Ridley Stroop do?

The original 1935 publication summarized three experiments completed by Stroop as his doctoral dissertation at George Peabody College (now part of Vanderbilt University). We'll focus on the first two experiments because they each illustrate a within-subjects design with one independent variable, tested at two levels, and using reverse counterbalancing. In the first experiment, 14 males and 56 females performed two tasks. Both involved reading the names of color words. Stroop (1992, p. 16) called one of the conditions RCNb ("**R**eading **C**olor **N**ames printed in **b**lack"). Participants read 100 color names (e.g., GREEN) printed in black ink as quickly and accurately as they could. The second condition Stroop (1992, p. 16) called RCNd ("**R**eading **C**olor **N**ames where the color of the print and the word are **d**ifferent"). In this case, the 100 color names were printed in colored ink, but the colors of the ink did not match the color name (e.g., the word GREEN was printed in red ink). The subjects' task was to read the word (e.g., the correct response is "green").

As a good researcher, Stroop (1935) was aware of the problems with *order effects*, so he used reverse counterbalancing (ABBA) to deal with the problem. After subdividing each of the stimulus lists into two sets of 50 items, Stroop gave some participants the sequence RCNb–RCNd–RCNd–RCNb, and an equal number of participants the sequence RCNd–RCNb–RCNb–RCNd. Thus, each subject read a total of 200 color names.

In Experiment 1, Stroop (1935) found *no difference* in performance between the RCNb and RCNd conditions. The average amount of time to read 100 words of each type was 41.0 seconds and 43.3 seconds, respectively. Reading the color names in the RCNd condition was unaffected by having the words printed in contrasting colors. It was in Experiment 2 that Stroop found the huge difference that eventually made his name so well known. Using the same basic design, this time the response was *naming the colors* rather than reading color names. In one condition, NC ("**N**aming **C**olor test"), participants named the colors of square color patches. In the second and key condition, NCWd ("**N**aming **C**olor of **W**ord test, where the color of the print and the word are **d**ifferent"), participants saw the same material as in the RCNd condition of Experiment 1, but this time, instead of reading the color name, they were to name the color in which the word was printed. If the letters of the word GREEN were printed in red ink, the correct response this time would be "red," not "green." Stroop's subjects had the same difficulty experienced by people trying this task today. Because reading is such an overlearned and automatic process, it interferes with the color naming, resulting in errors and slower reading times. Stroop found the average color naming times were 63.3 seconds for condition NC and a whopping 110.3 seconds for the NCWd condition. We've taken the four outcomes, reported by Stroop in the form of tables, and drawn a bar graph of them in Figure 7.2. As you can see, the Stroop effect is a robust phenomenon.

(continued)

BOX 7.1 (CONTINUED)

We mentioned earlier that Stroop actually completed three experiments for his dissertation. The third demonstrated that participants could improve on the NCWd task (the classic Stroop task) if given practice. An unusual aspect of this final study was that in the place of square color patches on the NC test, Stroop, for control purposes, substituted color patches in the shape of swastikas. This "made it possible to print the NC test in shades which more nearly match[ed] those in the NCWd test" (Stroop, 1992, p. 18). The choice probably reflected Stroop's religiosity—he regularly taught Bible classes and was well-known in the South for writing a series of instructional books called *God's Plan and Me* (MacLeod, 1991). He would have known the swastika originated as an ancient religious symbol, formed by bending the arms of a traditional Greek cross (+). Ironically, Stroop's study was published the same year (1935) the swastika was officially adopted as the symbol for the National Socialist (or "Nazi" for short) party in Germany.

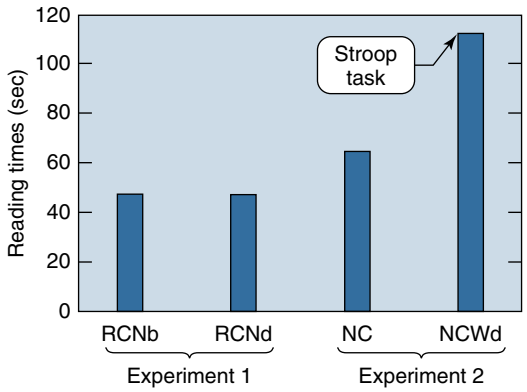

Figure 7.2
Combined data from the first two experiments of the original Stroop study (1935).

Two other counterbalancing strategies for a study with two conditions, when each condition is being tested many times, are simply to alternate the conditions (ABAB . . .) or to just present the sequences randomly. The latter approach was taken in the following experiment.

Research Example 14—Two-Level Repeated Measures Design

If you taste a nice piece of chocolate, do you think your experience will be influenced by having someone sitting next to you also eating the same kind of chocolate? Based on a long history of research in social psychology showing that people can influence others in a wide variety of ways, Boothby, Clark, and Bargh (2014) hypothesized that if two people share the same experience, their reactions to that experience might be amplified. They designed a clever within-subjects experiment with two conditions—shared experience and unshared experience. The experience to be measured involved tasting and then rating chocolate. A within-subjects design was chosen because this is a typical design in experiments involving a series of perceptual judgments.

The basic task was for participants (all female) to taste a piece of chocolate and then rate the experience along several dimensions (e.g., liking, how flavorful it was). They completed the task in the same room with a woman who appeared to be another subject, but was actually an experimental *confederate*, a term you recall from Chapter 3. In the shared experience condition, both women (participant and confederate) tasted and rated chocolate; in the unshared condition, the participant rated the chocolate while the confederate rated several paintings. In both conditions, the two "subjects" were told not to talk to each other.

You might suspect that subjects would be a little suspicious about having to rate chocolate two times in a row. In studies involving some level of deception, researchers always try to create a believable script, what they refer to as a *cover story*. In this case, subjects were led to believe they would be completing four tasks—rating two different varieties of chocolate and rating two different sets of paintings. In fact, the subjects only completed two tasks. They rated what they thought were two different types of chocolate (but were in fact from the same chocolate bar)—one with

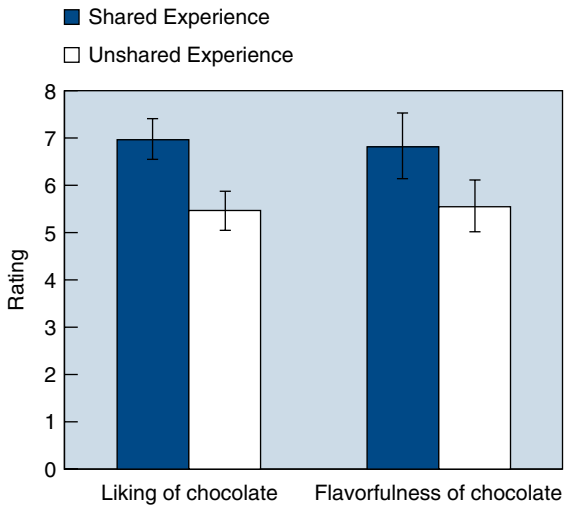

Figure 7.3
Judgments of how much chocolate was liked (left-hand bars) and how flavorful it seemed to be (right-hand bars) for those sharing or not sharing the chocolate-eating experience. From Boothby, Clark, and Bargh (2014).

the confederate also tasting chocolate (shared condition) and one with the confederate rating a booklet of artwork (unshared condition) with the counterbalanced order determined randomly. To enhance the cover story, the researchers also arranged for what seemed to be random drawings to determine who would be doing each task—the subjects believed they would always rate chocolates before the confederates would.

Figure 7.3 shows the results for the ratings of how much the subjects liked the chocolate and how flavorful they judged the chocolate to be. As you can see, in the shared experience condition, participants liked the chocolate better and judged it to have more flavor than in the unshared condition. Hence, sharing an experience seems to amplify the perception of that experience, even in the absence of any discussion about the experience.

Boothby et al.'s (2014) experiment involved a pleasant experience, and they wondered (*what's next thinking*) if the shared experiences truly intensified a perception or whether the sharing just made experiences more pleasant. As they put it, perhaps "Tootsie Rolls and fried tarantulas alike could be more palatable if tasted with another person" (p. 2212). This led them to replicate the study with bitter tasting chocolate. In this case the shared experience led subjects to a greater level of *dislike* for the chocolate; that is, sharing experiences intensifies whatever the initial experience, positive or negative, might be. The authors concluded that "every day, people spend time together in the absence of explicit communication. . . . Yet even in silence, people often share experiences, and the mental space inhabited together is a place where good experiences get better and bad experiences get worse" (p. 2215).

Self Test

7.1

1. There are two groups in a study, and participants are randomly assigned to one group or the other. What's the design?
2. Ellen signs up for a study in which she completes the Stroop test. Then she is asked to do the task again, but this time the words are turned upside down. What's the design?
3. What is the name of the design used when a subject variable is the independent variable?

Single Factor—More Than Two Levels

When experiments include a single independent variable, using two levels is the exception rather than the rule. Most single-factor studies use three or more levels and, for that reason, they are called **single-factor multilevel designs**. One distinct advantage of multilevel designs is they enable the researcher to discover **nonlinear effects**. To take a simple example, consider the well-known Yerkes-Dodson Law.

In their original research, Yerkes and Dodson (1908) taught mice a simple two-choice discrimination (choose to enter a chamber with white walls or black walls), shocking them for errors. In one of their experiments, they found that the mice learned faster as the shock intensity increased, but only up to a certain point; once shock reached high levels, performance declined. This simple finding has evolved over the years into a general "law" about arousal and performance, but the study involved a very small sample of mice, had a great deal of variability among those mice, and might have been analyzed using the wrong dependent variable (Teigen, 1994).[2] Figure 7.4 or one very much like it appears in a wide range of textbooks, showing that performance (athletic performance is the typical example) will be poor with low arousal, improves as arousal increases, and then declines when arousal becomes too high. Notice the value of having more than two levels of some "arousal" independent variable. If you only used low and moderate arousal (i.e., just two levels), you would conclude simply that performance increases as arousal increases. If you only used moderate and high arousal, you would conclude that performance decreases as arousal increases. And if you happened to use only very low and very high arousal, you would conclude that arousal does not affect performance at all. Using all three levels, resulting in the nonlinear effect, gives you a better overall picture of the effect of arousal – performance is best at moderate levels of arousal, and poor at either high or low levels of arousal.

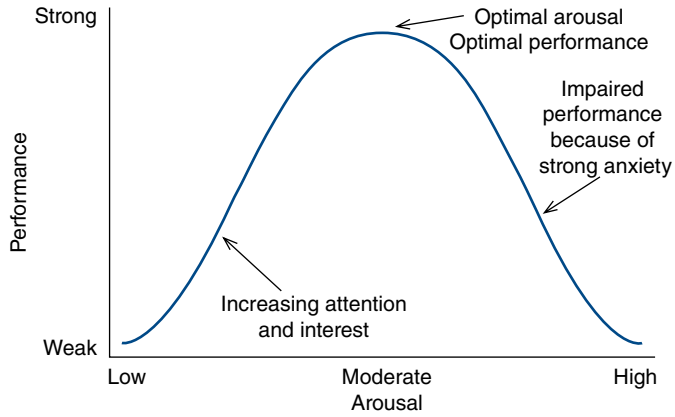

Figure 7.4
A typical presentation of the Yerkes–Dodson Law.

In addition to identifying nonlinear relationships, single-factor multilevel designs can also test for specific alternative hypotheses and perhaps rule them out while supporting the researcher's hypothesis. This is a strategy you will recognize from the discussions in Chapters 1 and 3 on the merits of *falsification* thinking. For example, from the literature on memory, consider a well-known series of studies by Bransford and Johnson (1972). In one of their procedures, subjects tried to memorize a paragraph that, in the absence of a topic, made little sense, as you can judge for yourself:

> The procedure is actually quite simple. First you arrange things into different groups. Of course, one pile may be sufficient depending on how much there is to do. If you

[2] Despite the apparent disconnect between the original study by Yerkes and Dodson and the subsequent "law" relating general arousal to motor performance, there is a long history of research showing that the Yerkes-Dodson principle works reasonably well, at least under certain circumstances (e.g., moderately difficult tasks).

> have to go somewhere else due to lack of facilities that is the next step, otherwise you are pretty well set. It is important not to overdo things. That is, it is better to do too few things at once than too many. In the short run this may not seem important but complications can easily arise. A mistake can be expensive as well. At first the whole procedure will seem complicated. Soon, however, it will become just another facet of life. It is difficult to foresee any end to the necessity for this task in the immediate future, but then one never can tell. After the procedure is completed one arranges the materials into different groups again. Then they can be put into their appropriate places. Eventually they will be used once more and the whole cycle will then have to be repeated. However, that is part of life. (p. 722).

In an *independent groups design*, Bransford and Johnson (1972) assigned participants to one of three groups. Those in Group 1 ("No Topic") had the paragraph read to them by the experimenter and then (after answering some questions about how well they understood the paragraph) tried to recall as much of it as they could. Those in Group 2 ("Topic Before") were told before they read the paragraph that "The paragraph you will hear will be about washing clothes" (p. 722). After the paragraph was read, and before recall occurred, group 3 participants ("Topic After") were told that "It may help you to know that the paragraph was about washing clothes" (p. 723). Bransford and Johnson identified 18 possible idea units that could be recalled and here are the average idea units recalled by each group[3]:

No Topic	$M = 2.82$	$SD = 2.47$
Topic Before	$M = 5.83$	$SD = 2.02$
Topic After	$M = 2.65$	$SD = 2.18$

Bransford and Johnson were interested in finding out if memory would be better if participants were giving an overall framework (or context) for the information being read to them. So they could have done a study with two levels, No Topic and Topic Before, and they would indeed have found evidence that providing a context ("the paragraph is about washing clothes") led to better memory performance (5.83 > 2.82). By adding the third level of the independent variable, however, they were able to evaluate if memory is better during the time when the information was being stored in memory (Topic Before) or at the time when the information was recalled from memory (Topic After). Their results showed them that providing a framework helps memory, but *only* if the framework is provided before the material to be learned. That is, they were able to *rule out* the hypothesis that providing a framework helps memory, regardless of when the framework is given.

Between-Subjects, Multilevel Designs

As with the two-level designs, the multilevel designs include both between- and within-subjects designs of the same four types: independent groups designs, matched groups designs, ex post facto designs, and within-subjects or repeated measures designs. Here is clever example of a between-subjects multilevel independent groups design.

Research Example 15—Multilevel Independent Groups Design

In Chapter 3 you read about the origins of Latane and Darley's (1968) groundbreaking research on helping behavior. Based on media reports of a real case (the murder of Kitty Genovese), they discovered and named the *bystander effect*—the tendency of people to fail to help if there are other bystanders witnessing the event. A considerable amount of subsequent research examined this effect and all the conditions enhancing or inhibiting it, but virtually no studies had used

[3] Bransford and Johnson (1972) reported standard error of the mean as their measure of variability within the sample. We converted standard errors to standard deviations.

young children as participants. Whether young children would show a bystander effect was the empirical question asked by Plötner, Over, Carpenter, and Tomasello (2015).

The experiment included three groups with random assignment used to place subjects into groups. In the "alone" condition, there was a single 5-year-old child in a room with a teacher who was in need of help. In the "bystander" condition, the child was joined by two other children of the same age, and in the "bystander-unavailable" condition, there were also two other children, but they were unable to help when the time came for help to be given. The situation requiring help was the teacher spilling a cup of colored water while painting at her desk, and helping was *operationally defined* as the child participant leaving his or her (half the participants were male, half female) chair and bringing the teacher some conveniently placed paper towels within 90 seconds. In the two bystander conditions, the other two children were 5-year-old *confederates*, and their use created a methodological issue seldom seen with confederates. Because of their age, they could not always be relied upon to stick to the script, so Plötner et al. (2015) had to eliminate some data due to "confederate error" (e.g., hinting about what was to happen). To be sure that trials for confederate error were not being deleted in some biased fashion, the researchers went as far as to calculate *inter-rater reliability* on the decision to exclude data, and reliability was high.

The results are pretty clear from Figure 7.5. When alone and when bystanders were unable to help (a barrier prevented them from getting to paper towels) almost all the children helped; when bystanders were present and able to help, however, the children helped just over half the time. Notice the advantage of going beyond just two IV levels and adding the third group. Had the study only included the alone and bystander groups, a bystander effect would have been demonstrated, but adding the bystander-unavailable condition allowed Plötner et al. (2015) to rule out two of the reasons typically given for the bystander effect (shyness, social referencing) while supporting a third reason (diffusion of responsibility, the idea that with several bystanders present,

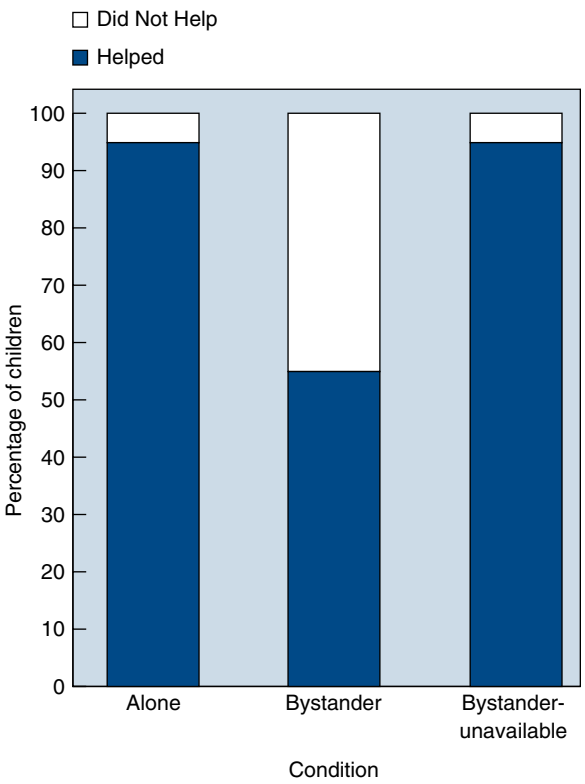

Figure 7.5
Percentage of children helping a teacher in a study of the bystander effect. From Plötner, Over, Carpenter, and Tomasello (2015).

the responsibility to help is shared and any one individual feels less responsible to help). As Plötner et al. described it,

> When bystanders were present but confined behind a barrier and therefore unavailable to help, children helped just as often as they did when they were alone. Thus, it was not simply the mere presence of bystanders that caused the effect (e.g., through shyness to act in front of others). Nor was it social referencing of the bystanders' passivity, as participants looked toward the bystanders equally often irrespective of their availability to help. . . . Rather, it appears that the effect was driven by diffusion of responsibility, which existed only in the bystander condition. (p. 504)

One final point about the study is that it could be considered an example of a *conceptual replication*. That is, Plötner et al. (2015) took a phenomenon (the bystander effect) that normally has been studied with adult subjects, and extended it to a different population, very young children.

Within-Subjects, Multilevel Designs

Whereas a single-factor, repeated-measures design with two levels has limited counterbalancing options, going beyond two levels makes all the counterbalancing options available. If each condition is tested once per subject, then both full and partial counterbalancing procedures are available. And when each condition is tested several times per subject, both reverse and blocked randomization procedures can be used. In the following study, each condition was tested just once, and partial counterbalancing (a Latin square) was used.

Research Example 16—Multilevel Repeated Measures Design

In the Chapter 3 discussion of replication, we briefly mentioned a study by Steele, Bass, and Crook (1999) that failed to replicate a controversial study by Rauscher, Shaw, and Key (1993). The Rauscher et al. study apparently discovered that small but significant improvements in spatial skills could follow from listening to music by Mozart. News of the research reached the popular press, which often distorts accounts of scientific research, and soon stories appeared urging parents to play Mozart to their infants to improve their cognitive abilities. Yet as we saw in Chapter 3, several replication efforts failed. Another study by Steele and his colleagues further questioned the viability of this alleged "Mozart effect."

Steele, Ball, and Runk (1997) included three conditions in their experiment: listening to Mozart for 10 minutes, listening to a recording of soothing environmental sounds (a gentle rainstorm) for 10 minutes, and not listening to anything (sitting quietly for 10 minutes and trying to relax). All 36 participants in the study were tested in each of the three conditions, making this a single-factor, multilevel, repeated-measures design. Although complete counterbalancing would have been easy to implement ($3! = 3 \times 2 \times 1$, or six orders of conditions, with six subjects randomly assigned to each order), the authors chose to use a 3×3 *Latin square*, with 12 participants randomly assigned to each row of the square. To avoid the bias that might result if participants thought they were evaluating the Mozart effect, the study's *cover story* was that "the experiment concerned the effect of relaxation on recall" (Steele et al., 1997, p. 1181). Instead of using a spatial skills task, Steele et al., used a difficult memory task, a backward digit span procedure where subjects must repeat the sequence of digits in backward order from which it was provided. For example, given a stimulus such as "6-8-3-1-7," the correct response would be "7-1-3-8-6." On a given trial, participants would listen to Mozart, listen to gentle rainfall, or sit quietly, and then be given three consecutive digit span trials. Each digit span included nine numbers, presented in a random order. Thus, a score from 0 to 27 could be earned for the three trials combined.

The study produced statistically significant findings, but none that would comfort those advocating for the Mozart effect. The average number of digits correctly recalled was virtually identical for all three conditions: 18.53 (*SD* = 4.14) for Mozart, 18.50 (*SD* = 6.07) for the gentle rain, and 18.72 (*SD* = 5.09) for the control condition. There was a significant practice effect, however. Regardless of the order in which the conditions were presented, participants improved from the first set of digit span tests to the third set. From the first to the third, the averages were 15.64, 19.14, and 20.97 (*SD*s of 4.70, 4.87, and 4.29, respectively). So, should parents play Mozart tapes for their children? Sure, why not? Will it make them smarter? Apparently not, although it could make them enjoy classical music, an outcome of value by itself.

Analyzing Data From Single-Factor Designs

As we described in Chapter 4, researchers use both descriptive and inferential statistics to evaluate data from their studies. In this section, we discuss ways to visually depict descriptive data, then we turn to inferential analysis of data.

Presenting the Data

One decision to be made when reporting the results of any research study is how to present the data. There are three choices. First, the numbers can be presented in sentence form, an approach that might be fine for reporting the results of experimental studies with two or three levels (e.g., the Mozart example) but makes for tedious reading as the amount of data increases. A second approach is to construct a table of results. Usually, means and standard deviations for each condition are presented in tables. A table for the Bransford and Johnson (1972) study, using APA format, would look like Table 7.2.

A third way to present the data is in the form of a graph. Note that in an experimental study (e.g., Figure 7.5), a graph always places the dependent variable on the vertical (*Y*) axis and the independent variable on the horizontal (*X*) axis. The situation becomes a bit more complicated when more than one independent variable is used, as you will see in the next chapter. Regardless of the number of independent variables, the dependent variable *always* goes on the vertical axis.

Deciding between tables and figures is often a matter of the researcher's preference. Graphs can be especially striking if there are large differences to report or (especially) if nonlinear effects (e.g., the Yerkes-Dodson Law) occur or if the result is an *interaction* between two factors (coming in Chapter 8). Tables are often preferred when data points are so numerous that a graph would be uninterpretable or when the researcher wishes to inform the reader of the precise values of the means and standard deviations. One rule you can certainly apply is to never present the same data in both table and graph form. In general, you should present data in such a way that the results you have worked so hard to obtain are shown most clearly.

Table 7.2 Bransford and Johnson's (1972) Data in Table Format Mean Number of Idea Units Recalled as a Function of Different Learning and Recall Contexts

Condition	Mean Score	Standard Deviation
No Topic	2.82	2.47
Topic Before	5.83	2.02
Topic After	2.65	2.18

Note. The maximum score was 18.

Types of Graphs

When we first described the data from Bransford and Johnson's (1972) study on memory for doing laundry, we simply listed the means and standard deviations for the three groups. An alternative would have been to present the data in the form of a bar graph, as in Figure 7.6.

The graph clearly shows the advantage of context on memory. But could we also have shown these data in the form of a line graph, as in Figure 7.7?

The answer is no. The problem concerns the nature of the construct used as the independent variable and whether the independent variable is a between-subjects factor or a within-subjects factor. If the independent variable is manipulated between-subjects or is a subject variable, then the type of graph to use is a bar graph. (Remember B: between-subjects = bar graph). The reason for this is because the levels of the independent variable represent separate groups of individuals, so the data in the graph should best reflect separate groups, or separate bars in this case.[4] The top

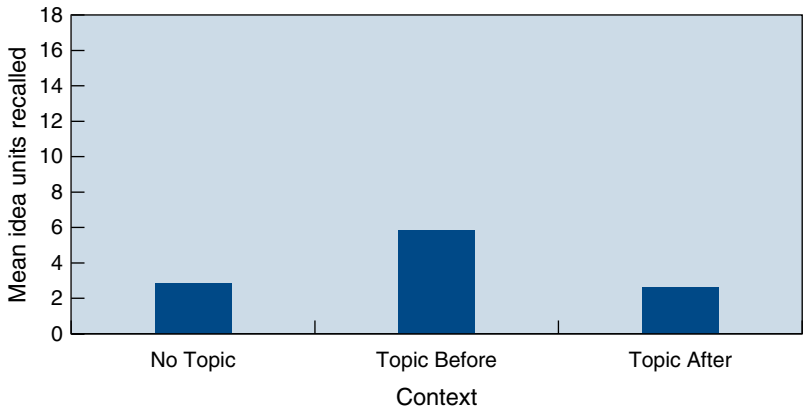

Figure 7.6
The Bransford and Johnson (1972) "laundry" study presented as a bar graph.

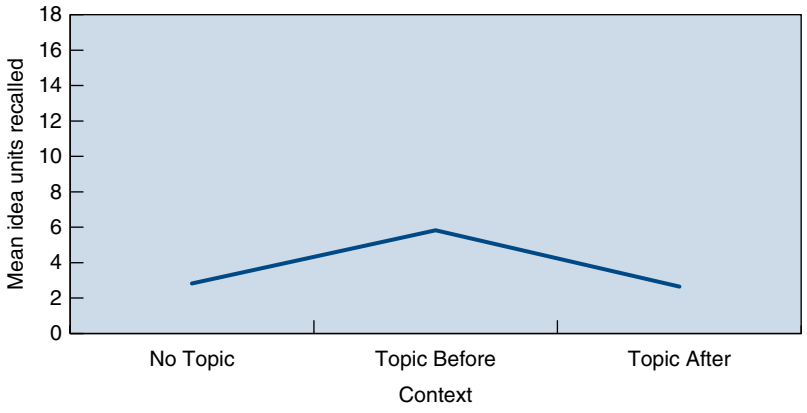

Figure 7.7
The Bransford and Johnson (1972) "laundry" study presented inappropriately as a line graph.

[4] Generally speaking, bar graphs are used with between-subjects designs, but if the independent variable represents a continuous-type variable, then a line graph can be used. For example, in a study that manipulates drug dosage levels (3 mg, 5 mg, and 7 mg) between different groups of rats, the dosage level is a continuous variable, in which the variable exists on a continuum. In such cases, the researcher may opt to represent the means of each dosage level as points on a continuous line (i.e., a line graph).

of each bar represents the mean for each condition in the study. Often, researchers will also place *error bars* on the tops of the graphs which can reflect standard deviations or *confidence intervals* for each condition. If the independent variable is a within-subjects manipulated or subject variable, then it is appropriate to use a line graph. The reason for this is because participants are experiencing all levels of that independent variable, so the data should "connect" in a more continuous way. Thus, a line graph shows participants going through the levels of the independent variable. On a line graph, the points usually represent the means of each condition, and error bars are typically placed on each point on the graph.

In general then, bar graphs can be used for between-subjects designs, and line graphs should be used for within-subjects designs. Be sure to review Box 4.3 in Chapter 4 for a reminder about the ethics of presenting the data. It is easy to mislead the uninformed consumer of research by, for example, altering the distances on the *Y*-axis. Figure 7.8 is a "gee whiz" graph illustrating Bransford and Johnson's (1972) data after altering the *Y*-axis.

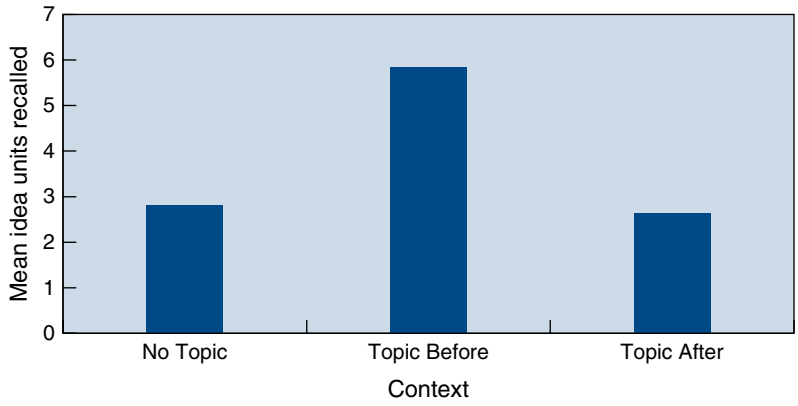

Figure 7.8
The Bransford and Johnson (1972) "laundry" study presented inappropriately as "Gee whiz!" graph.

It appears that the maximum score one could get on the recall test was 7, but really it was 18. Your responsibility as a researcher is to present your results honestly and in the way that best illustrates the true outcomes of your study.

One of psychology's most famous line graphs continues to appear in introductory psychology texts even though (a) the study was completed more than 130 years ago and (b) the study's author never presented that data in graph form. For more on the study that is said to have originated experimental research on memory, read Box 7.2.

BOX 7.2 ORIGINS—The Ebbinghaus Forgetting Curve

Trained as a philosopher and interested in the ancient philosophical problem of how ideas become associated together in the mind, Hermann Ebbinghaus (1850–1909) completed the first systematic study of memory in psychology's history. The work was summarized in his brief (123 pages in a 1964 reprinting) *Memory: A Contribution to Experimental Psychology* (Ebbinghaus, 1885/1964).

The first task for Ebbinghaus was to find materials that would not normally be associated with each other. His solution, considered to be a classic example of scientific creativity, was to create about 2300 stimuli, each consisting of a consonant, then a vowel, and then a second consonant (e.g., ZEK, KIG). These "CVCs" (or "nonsense syllables" as they later came to be called) were not necessarily meaningless in themselves, but sequences of them were unlikely to bring old associations to mind, Hence, memorizing a list of them would constitute, as far as Ebbinghaus was concerned, the creation of a brand new set of associations. For several

years, showing great perseverance and/or a complete lack of social life, Ebbinghaus spent several hours a day memorizing and recalling lists of CVCs, Yes, he was the *only* subject in the research. He carefully studied such factors as the number of CVCs per list, the number of study trials per list, and whether the study trials were crammed together or spread out.

Ebbinghaus's most famous study examined the time course of forgetting; his empirical question was "Once some material has been memorized, how much of that memory persists after varying amounts of time?" His outcome illustrates two ideas you have learned about in this chapter—the appearance of nonlinear effects when there are more than two levels of an independent variable, and the effective use of a line graph with a within-subjects, manipulated variable (time). The famous Ebbinghaus forgetting curve is shown in Figure 7.9. Although Ebbinghaus presented his data in table form, most subsequent descriptions of the study have used a line graph like the one shown here.

In his study of forgetting, Ebbinghaus memorized lists of 13 CVCs and then tried to recall them after various amounts of time—from the *X*-axis, you can see that the levels of his independent variable (retention interval) were 20 minutes, 1 hour, 8.8 hours, 1 day, 2 days, 6 days, and 31 days. His dependent variable used what he called the method of savings. He would learn a list, wait for the retention interval to pass, and then try to relearn the list again. If it took him 20 minutes to learn the list at first, and 5 minutes to relearn it, he determined that he had "saved" 15 minutes. His *Y*-axis, "% saved" was calculated this way: [(original learning time–relearning time)/original learning time]. For this example, the savings would be 75%. Ebbinghaus's method of savings has an interesting implication for you as a student. You might think that you don't remember anything about your introductory psychology course, but the fact that you would relearn the material more quickly than you learned it originally means that some memory is still there to help you in your relearning of the material.

As you can see from the graph, recall declined as the retention interval increased, but the decline was not a steady or linear one. Instead, a nonlinear effect occurred. Forgetting occurred very rapidly at first, but then the rate of forgetting slowed. Thus, after a mere 20 minutes, only about 60% (58.2 actually) of the original learning had been saved. At the other end of the curve, there wasn't much difference between an interval of a week (25.4% saved) and a month (21.1% saved).

From the standpoint of methodological control, there are several other interesting points about the Ebbinghaus research. To ensure a constant presentation rate, for example, he set a metronome to 150 beats per minute and read each CVC on one of the beats. He also tried to study the lists in the same environment and at about the same time of day, and to use no memorization technique except simple repetition. Also, Ebbinghaus worked only when sufficiently motivated so that he could "keep the attention concentrated on the tiresome task" (Ebbinghaus, 1885/1964, p. 25). Finally, he understood the importance of *replication*. He completed one set of studies in 1879–1880 and then replicated all his work three years later, in 1883–1884.

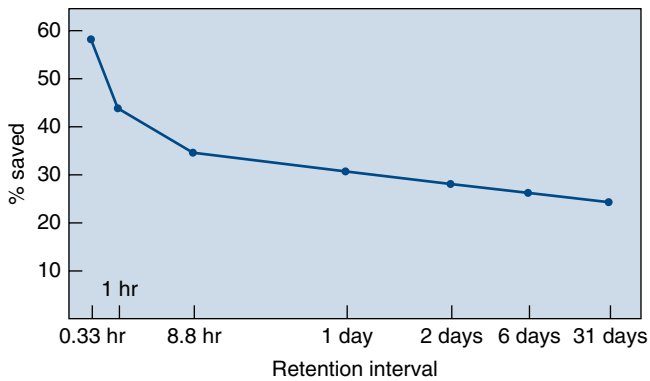

Figure 7.9
The Ebbinghaus forgetting curve—appropriately drawn as a line graph and a good illustration of a nonlinear effect.

Analyzing the Data

To determine whether the differences found between the conditions of a single-factor design are significant or due to chance, inferential statistical analysis is required. An inferential statistical decision to reject or not reject the null hypothesis in a study depends on analyzing two types of variability in the data. The first refers to the differences between groups, which are caused by a combination of (a) systematic variance and (b) error variance. **Systematic variance** is the result of an identifiable factor, either the variable of interest or some factor that you've failed to control adequately (such as confounds). **Error variance** is nonsystematic variability due to individual differences between subjects in the groups and any number of random, unpredictable effects that might have occurred during the study. Error variance also occurs within each group, also as a result of individual differences and other random effects, and accounts for the differences found there. Mathematically, many inferential analyses calculate a ratio that takes this form:

$$\text{Inferential statistic} = \frac{\text{Variability between conditions}\,(\text{systematic} + \text{error})}{\text{Variability within each condition}\,(\text{error})}$$

The ideal outcome is to find that variability between conditions is large and variability within each condition is small. An inferential statistic will then test to see if the researcher can reject the null hypothesis and conclude with a certain degree of confidence that there is a significant difference between the levels of the independent variable.

Recall from Chapter 4 that *inferential statistics* are used to infer from a sample what might occur in the population of interest. Inferential statistical tests can be either parametric tests or nonparametric tests. Parametric tests are tests which have certain assumptions (or parameters) that are required to best estimate the population. For example, some tests assume your data in level of the independent variable approximates a normal distribution. For this reason, it is important to examine the distributions of data so you can select the appropriate statistical test. Another assumption is called **homogeneity of variance**, which means the variability of each set of scores being compared ought to be similar. So, if the standard deviation for one group is significantly larger than the standard deviation for the other group, there may be a violation of the assumption of homogeneity of variance. Tests for homogeneity of variance exist, and if these tests indicate a violation, nonparametric tests can be used. These tests that do not have the same assumptions as parametric tests and can be used if violations of the parameters for your ideal statistical tests occur.

Scales of measurement are also important to consider when selecting the appropriate statistical test for a particular type of research design. For single-factor designs when interval or ratio scales of measurement are used (and the aforementioned parameters are met), then *t*-tests or one-way ANOVA (ANalysis Of VAriance) are calculated. Other techniques are required when nominal or ordinal scales of measurement are used. For example, a chi-square test of independence could be used with nominal data.

Statistics for Single-Factor, Two-Level Designs

There are two varieties of the *t*-test for comparing two sets of scores. The first is called an **independent samples *t*-test**, and, as the name implies, it is used when the two groups of participants are completely independent of each other. This occurs whenever we use random assignment to create equivalent groups, or if the variable being studied is a subject variable involving two different groups (e.g., males versus females). If the independent variable is a within-subjects factor, or if two groups of people are formed in such a way that some relationship exists between them (e.g., participants in Group A are matched on intelligence with participants in Group B), a

dependent samples or paired *t*-test is used. For the four single-factor designs just considered, the following *t*-tests would be appropriate:

- independent samples *t*-test
  - independent groups design
  - ex post facto design
- dependent samples *t*-test
  - matched groups design[5]
  - repeated-measures design

In essence, the *t*-test examines the difference between the mean scores for the two samples and determines (with some probability) whether this difference is larger than would be expected by chance factors alone. If the difference is indeed sufficiently large, and if potential confounds can be ruled out, then the researcher can conclude with a high probability that the differences between the means reflect a real effect. Be mindful, though, that the *t*-test is a parametric statistic with the assumptions of normal distributions of data and homogeneity of variance. So, if your data don't meet these assumptions, then an alternate nonparametric test can be used, such as the Mann-Whitney *U*-test.

As you recall from Chapter 4, in addition to determining if differences are statistically significant, it is also possible to determine the magnitude of the difference by calculating *effect size*, usually Cohen's *d* for two-sample tests. To learn (or, I hope, review) the exact procedures involved, both in calculating the two forms of *t*-test and in determining effect size, please review the Student Statistics Guide on the Student Companion site or consult any introductory statistics text (e.g., Witte & Witte, 2014). Also in that guide, if your school uses SPSS (a common statistical software package), you can use learn how to perform both types of *t*-tests in SPSS.

Statistics for Single-Factor, Two-Level Designs

For single-factor, multilevel designs such as Bransford and Johnson's (1972) laundry study, you might think the analysis would be a simple matter of completing a series of *t*-tests between all of the possible pairs of conditions (i.e., No Topic versus Topic Before; No Topic versus Topic After; Topic Before versus Topic After). Unfortunately, matters aren't quite that simple. The difficulty is that completing multiple *t*-tests increases the risks of making a Type I error—that is, the more *t*-tests you calculate, the greater the chances that one of them will accidentally yield significant differences between conditions. In a study with five levels of the independent variable, for example, you would have to complete 10 *t*-tests to cover all of the pairs of levels of the independent variable.

The chances of making at least one Type I error when doing multiple *t*-tests can be estimated by using this formula:

$$1-\left(1-\text{alpha}\right)^{c}$$

where c = the number of comparisons being made.

[5] You may be wondering why a matched group design, which involves random assignment of participants to separate groups, would require a dependent samples *t*-test. The reason is that before random assignment to condition, participants are matched on at least one variable, and conceptually, this makes the individuals that are paired together with similar scores essentially one "person." Thus, when the pair is split and participants with similar scores are assigned to different, independent groups, you essentially have the same "person" in each level of the independent variable. Therefore, you are treating the data like within-subjects data, which requires (in this case) a dependent samples *t*-test.

Thus, if all the possible *t*-tests are completed in a study with five levels, there is a very good chance (4 out of 10) of making at least one Type I error:

$$1-(1-0.05)^{10} = 1-(0.95)^{10} = 1-0.60 = 0.40 \text{ or } 40\%$$

To avoid the problem of multiple *t*-tests in single-factor designs, researchers typically use a procedure called a *one-way* (sometimes *one-factor*) *analysis of variance* or one-way **ANOVA** (**AN**alysis **O**f **VA**riance). The *one* in the "one-way" means one independent variable. In essence, a one-way ANOVA tests for the presence of an overall significant effect that could exist somewhere between the levels of the independent variable. Hence, in a study with three levels, the null hypothesis is "level 1 = level 2 = level 3." Rejecting the null hypothesis does not identify which condition differs from which, however. To determine precisely which condition is significantly different from another requires subsequent testing or post hoc (after the fact) analyses. In a study with three levels, subsequent testing would analyze each of the three pairs of comparisons, but only after the overall ANOVA has indicated some significance exists. Selecting one of the types of post hoc analyses depends on sample size and how conservative the researcher wishes to be when testing for differences between conditions. For example, Tukey's HSD test, with the HSD standing for "honestly significant difference," is one of the most popular of the post hoc choices (Sprinthall, 2000), but it requires that there are equal numbers of participants in each level of the independent variable. A more conservative test for comparisons of groups with unequal sample sizes per condition is the Bonferroni correction. SPSS provides many options for different post hoc tests, or you can consult a statistics textbook for more information. Importantly, if the ANOVA does not find any significance, subsequent testing is normally not done, unless specific predictions about particular pairs of levels of the independent variable have been made ahead of time. In this latter case, the testing is not post hoc tests, but referred to as *planned comparisons*.

The one-way ANOVA yields an *F* score or an *F* ratio. Like the calculated outcome of a *t*-test, the *F* ratio examines the extent to which the obtained mean differences could be due to chance or are the result of some other factor (presumably the independent variable). For a one-way ANOVA the inferential statistic is the *F* ratio. It is typically portrayed in a table called an **ANOVA source table**. An example of how one of these could be constructed for a one-way ANOVA for an independent groups design with three levels of the independent variable is in the Student Statistics Guide on the Student Companion Site. The complete calculations for this analysis, as well as a follow-up analysis for effect size, can also be practiced there, and the guide also shows you how to use SPSS to complete a one-way ANOVA as well as a Tukey's HSD test.

Recall that the independent samples *t*-test is used with independent groups and ex post facto designs, and the dependent samples *t*-test is used with matched groups and repeated measures designs. The same thing occurs for the one-way ANOVA. In addition to the one-way ANOVA for independent groups, there is also a one-way ANOVA for repeated measures. Parallel to the *t*-tests, these ANOVAs are used in these situations:

- one-way ANOVA for independent groups
  - multilevel independent groups design
  - multilevel ex post facto design
- one-way ANOVA for repeated measures
  - multilevel matched groups design
  - multilevel repeated-measures design

Again, be mindful of the parameters required for the use of the one-way ANOVA, which like the *t*-test, requires normal distributions of data and homogeneity of variance. If either or both are violated, alternate nonparametric tests should be used.

Self Test 7.2

1. Why must a study like the Bransford and Johnson study (effect of context on memory) be portrayed with a bar graph and not a line graph?
2. Suppose a researcher wanted to test the difference between women and men on the number of items answered correctly on a cognitive reasoning task. What statistical test(s) would be appropriate to use for this design?
3. Suppose a researcher wanted to test children at the beginning, middle, and end of the school year to see their progress on standardized math tests. What statistical test(s) would be appropriate to use for this design?

Special-Purpose Control Group Designs

We introduced the basic distinction between experimental groups and control groups in Chapter 5. As you recall, although control groups are not always needed, they are especially useful when the research calls for a comparison of a treatment of some kind (e.g., a drug effect) with a baseline level of behavior. Experimental groups receive the treatment, while those in the control group do not. In repeated-measures designs, a parallel distinction can be made between experimental conditions and control *conditions*, with both conditions experienced by each of the study's participants. Besides the typical control group situation in which a group is untreated, three other special-purpose control group designs are worth describing: placebo controls, wait list controls, and yoked controls. These types of control groups are most informative when used in the context of multilevel experimental designs.

Placebo Control Group Designs

A **placebo** (from Latin, meaning "I shall please") is a substance or treatment given to a participant in a form suggesting a specific effect when, in fact, the substance or treatment has no genuine effect. In drug research, for example, patients will sometimes show improvement when given a placebo but told it is a real drug, simply because they *believe* the drug will make them better. In research, members of a **placebo control group** are led to believe they are receiving a particular treatment when, in fact, they aren't. Can you see why this would be necessary? Suppose you wished to determine if alcohol slows reaction time. If you used a simple experimental group that was given alcohol and a second group that received nothing to drink, then gave both groups a reaction time test, the reactions indeed might be slower for the first group. Can you conclude that alcohol slows reaction time? No—participants might hold the general belief that alcohol will slow them down, and their reactions might be influenced by that knowledge. To solve the problem, you must include a group given a drink that seems to be alcoholic (and cannot be distinguished in taste from the true alcoholic drink) but is not. This group is the placebo control group. Should you eliminate the straight control group (no drinks at all)? Probably not, for these individuals yield a simple baseline measure of reaction time. If you include all three groups and get these average reaction times:

Experimental group:	0.32 second
Placebo control:	0.22 second
Straight control:	0.16 second

You could conclude that what people expect about the effects of alcohol slowed reaction time somewhat (from 0.16 to 0.22) but that alcohol by itself also had an effect beyond people's expectations (0.22 to 0.32). By the way, in terms of the designs introduced earlier in this chapter, this study would be an independent groups, single-factor, multilevel design, assuming subjects would be randomly assigned to groups. If subjects were first matched on some variable (e.g., matched for weight), the study would be a matched groups, single-factor, multilevel design. Can you think why an ex post facto design or a repeated-measures design would not be appropriate here?

Wait List Control Group Designs

Wait list control groups are often used in research designed to assess the effectiveness of a program (Chapter 11) or in studies on the effects of psychotherapy. In this design, the participants in the experimental group are in a program because they are experiencing a problem the program is designed to alleviate; wait list controls are also experiencing the problem. For instance, a study by Miller and DiPilato (1983) evaluated the effectiveness of two forms of therapy (relaxation and desensitization) to treat clients who suffered from nightmares. They wanted to include a no-treatment control, but to ensure clients in all three groups (relaxation, desensitization, and control) were generally equivalent, the control group subjects also had to be nightmare sufferers. From an identified pool of nightmare sufferers, participants were randomly assigned to one of the three groups, making this an independent groups, single-factor, multilevel design. For ethical reasons, those assigned to the wait list were assured they would be helped, and after the study ended they were given treatment equivalent to that experienced by the experimental groups.

Giving the wait list participants an opportunity to benefit from some therapy procedure provides an important protection for subject welfare, but it also creates pressures on the researcher to use this control procedure only for therapies or programs of relatively brief duration. In Miller and DiPilato's (1983) study, for example, the subjects in the two experimental groups were in relaxation or desensitization therapy for 15 weeks, and both forms of therapy produced a reduction in nightmares compared to the wait list control subjects. At the end of 15 weeks, those in the wait list control group began treatment (randomly assigned to either relaxation or desensitization, because both procedures had worked equally well).

Some might argue it is unethical to put people into a wait list control group because they won't receive the program's benefits right away and might be harmed while waiting. This issue can be especially problematic when research evaluates life-influencing programs. Read Box 7.3 for an examination of this issue and a defense of the use of control groups, including wait lists, in research.

BOX 7.3 ETHICS—Who's in the Control Group?

In a study on human memory in which an experimental group gets special instructions to use visual imagery, while a control group is told to learn the word lists any way they can, the question of who is assigned to the control group does not create an ethical dilemma. However, things are not so simple when an experiment is designed to evaluate a program or treatment that, if effective, would clearly benefit people, perhaps even by prolonging their lives. For example, in a well-known study of the effects of personal control on health (Langer & Rodin, 1976), some nursing home residents were given increased control over their daily planning, while control group residents had their daily planning done for them (for the most part) by the nursing staff. On the average, residents in the first group were healthier, mentally and physically, and were more likely to be alive when the authors came back and did an 18-month follow-up study (Rodin & Langer, 1977). If you discovered one of your relatives (now dead) had been assigned to the control group, do you think you would be upset?

In a similar vein, there has been controversy over the assignment of participants to control groups in studies with cancer patients (Adler, 1992). The research concerned the effects of support groups on the psychological well-being and physical health of women with breast cancer. The findings indicated that women in support groups recovered more quickly and even lived longer than women not placed in these groups (i.e., those in the control group). Some researchers argued the results did not reflect the benefits of support groups as much as the harm done to those in the control group who might feel left out or rejected. This could create stress, and it is known that stress can harm the immune system, leading to a host of health-related problems. So is there some truth to Figure 7.10? At the extreme, can being in a control group kill you?

Defenders of the control group approach to evaluating programs make three strong arguments. First, they point out that hindsight is usually perfect. It is easy to say after the fact that "a program as effective as this one ought to be available to everyone." The problem is that *before* the fact, it is not so obvious that a program will be effective. The only way to tell is to do the study. Prior to Langer and Rodin's (1977) nursing home study, for example, one easily could have predicted that the experimental group subjects would be unnecessarily stressed by the added responsibility of caring for themselves and drop like flies. Similarly, those defending the cancer studies point out that, when these studies began, few women expressed any preference about their assignment to either an experimental or a control group, and some actually preferred to avoid the support groups (Adler, 1992). Hence, it was not necessarily the case that control group participants would feel left out or overly stressed.

Second, researchers point out that in research evaluating a new treatment or program, the comparison is not between the new treatment and no treatment; it is between the new treatment and the most favored *current* treatment. So, for control group members, available services are not being withheld; they are receiving normal, well-established services. Furthermore, once the study has demonstrated a positive effect of the experimental treatment, members of the control groups are typically given the opportunity to be treated with the new approach.

Third, treatments cost money, and it is certainly worthwhile to spend the bucks on the best treatment. That cannot be determined without well-designed research on program effectiveness, however. In the long run, programs with empirically demonstrated effectiveness serve the general good and may save or prolong lives.

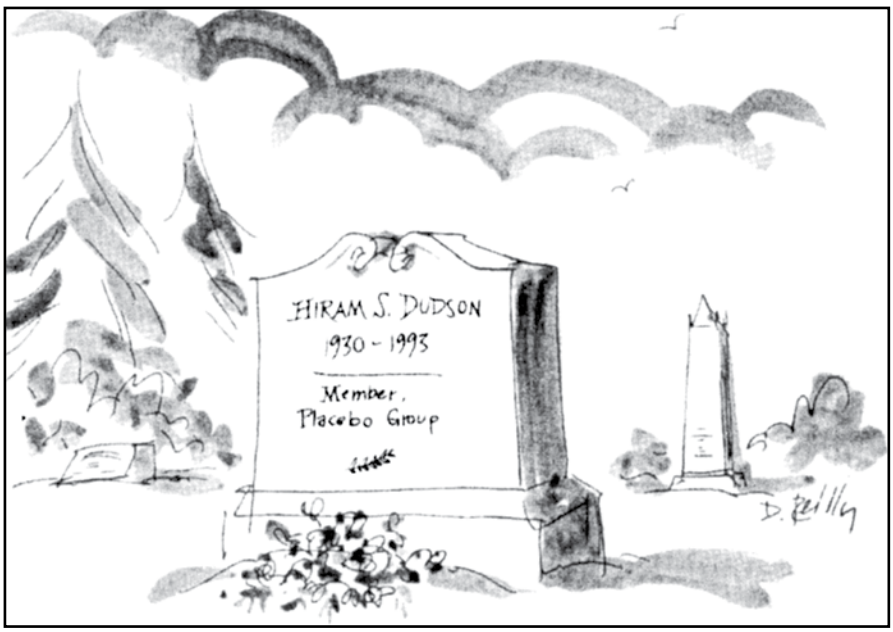

Figure 7.10
Potential consequences of being assigned to the control group.

Remember the Chapter 1 discussion of pseudoscience as illustrated by handwriting analysis. Another common example of pseudoscience involves the use of the so-called subliminal self-help. The idea is that while you listen to what appears to be the soothing sounds of ocean waves, subliminal messages (i.e., below the threshold of normal hearing) are sent that will be detected by your unconscious mind and lead you to make changes that will improve your life in some fashion. In fact, there is ample research indicating that any positive effects of these tapes are the result of what people *expect* to happen. Testing expectancy often involves the use of placebo controls, but consider the following Research Example, which effectively combines both placebos and wait lists to yield a different interpretation of what these self-help programs accomplish.

Research Example 17—Using Both Placebo and Wait List Control Groups

One of the favorite markets for the subliminal self-help business is in the area of weight loss. Americans in particular try to lose weight by attempting an unending variety of techniques, from fad diets to surgery. People are especially willing to try something when minimal effort is involved, and this is a defining feature of the subliminal approach; just open a file on your iPhone and pretty soon your unconscious will be directing your behavior so that weight loss will be inevitable. In a study that creatively combined a placebo control and a wait list control, Merikle and Skanes (1992) evaluated the effectiveness of self-help weight loss audiotapes (no iPhones in 1992). Forty-seven adult females were recruited through newspaper ads and randomly assigned to one of three groups. The experimental group participants (n = 15) were given a commercially produced subliminal self-help audiotape that was supposed to help listeners lose weight. Those in the placebo control group (n = 15) thought they were getting a subliminal tape designed for weight loss, but in fact they were given one designed to relieve dental anxiety (the researchers had a sense of humor). Based on *pilot study* results, the two tapes were indistinguishable to ordinary listeners. A third group, the wait list control (n = 17), was told "that the maximum number of subjects was currently participating in the study and that. . .they had to be placed on a waiting list" (p. 774). Those in the experimental and placebo groups were told to listen to their tapes for one to three hours per day and participants in all three groups were weighed weekly for five weeks.

The results? As you can see in Figure 7.11, those in the experimental group lost a modest amount of weight (very modest—check out the *Y*-axis), but the *same* amount was also lost by those in the placebo control group. This is the outcome typical with this type of study, indicating

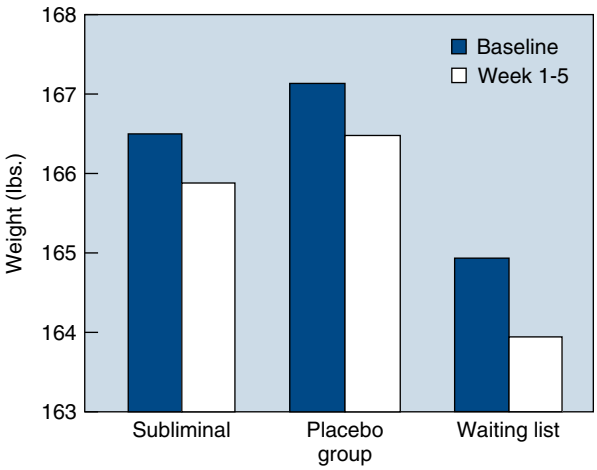

Figure 7.11
Results of the Merikle and Skanes (1992) study using placebo and wait list control groups to evaluate a subliminal self-help program for weight loss.

the subliminal tapes have no effect by themselves. The interesting outcome, however, was that the wait list group also lost weight—about the same amount as the other two groups. This result led Merikle and Skanes to conclude that subliminal tapes do not produce their results simply because of a placebo effect. If this had been true, the placebo group participants, believing their mind was being altered, would have lost more weight than the wait list group folks, who, not yet in possession of the tapes, would not have lost any weight. That subjects in all three groups lost weight led the authors to argue that simply being in an experiment on weight led subjects in all the groups to think about the problem they were experiencing. Subjects in the three groups "may have lost weight simply because participation in the study increased the likelihood that they would attend to and think about weight-related issues during the course of the study" (p. 776).

In this study, the wait list control group had the effect of evaluating the strength of the placebo effect and providing an alternative explanation for the apparent success of subliminal tapes. Also, although the authors didn't mention the term, the study's outcome sounds suspiciously like a *Hawthorne effect*, which you learned about in Chapter 6. And you might have noticed the rationale for adding the wait list control group is another example, like Bransford and Johnson's (1972) memory study described earlier, of adding more than two levels to an independent variable in order to directly test (and potentially rule out) certain hypotheses. The Merikle and Skanes study raises serious questions about the placebo effect hypothesis as an explanation for the effects of subliminal self-help recordings. It also provides additional evidence that there is no validity to the claim that subliminal recordings work.

One final point worth mentioning about this study is that the second author, Heather Skanes, was the experimenter throughout the study, and the first author, Philip Merikle, arranged the subliminal tape labels. Thus, he was the only one who knew who was getting the weight loss tape (experimental group) and who was getting the dental anxiety tape (placebo group). The authors thus built a nice *double blind* control into their study.

Yoked Control Group Designs

A third type of control group is the **yoked control group**. It is used when each subject in the experimental group, for one reason or another, participates for varying amounts of time or is subjected to different types of events in the study. Each member of the control group is then matched, or "yoked," to a member of the experimental group so that, for the groups as a whole, the time spent participating or the types of events encountered is kept constant. A specific example will clarify.

Research Example 18—A Yoked Control Group

A nice example of a yoked control group is a study by Dunn, Schwartz, Hatfield, and Wiegele (1996). The study was designed to evaluate the effectiveness of a psychotherapy technique that was popular (but controversial) in the 1990s. The therapy is called "eye movement desensitization and reprocessing," or EMDR. It is said to be effective as a treatment for anxiety disorders, especially posttraumatic stress disorder. The essence of the therapy is that the client brings to mind and concentrates on a personal traumatic event. While thinking about this event, the client follows a series of hand movements made by the therapist by moving the eyes rapidly from side to side. During the session, the client continuously rates the level of stress being experienced, and when it reaches a certain low point, the eye movement tracking stops. This might sound a bit fishy to you, as it did to Dunn and his colleagues. They wondered if a placebo effect might be operating—clients think the procedure will work and their expectations and faith in the therapist make them feel better. Most of the support for EMDR has been anecdotal (and you know from Chapter 1 to be skeptical about testimonials), so Dunn decided to use a stronger experimental test.

Dunn et al. (1996) identified 28 college students who had experienced mildly traumatic events (those found with more serious trauma were referred to the university counseling center), and after using a matching procedure to create equivalent groups (matching them for age, gender, and the type of traumatic event they reported), randomly assigned them to an experimental and a yoked control group. In the experimental group, participants underwent EMDR. As they thought about their traumatic event and tracked the experimenter's finger with their eyes, they periodically reported their level of stress on a 10-point "SUD" (Subjective Units of Discomfort) scale. Some physiological measures (e.g., pulse) were also recorded. This procedure continued until they reached a SUD level of 0–1 or until 45 minutes had elapsed. Hence, the therapy lasted for varying amounts of time for those in the experimental group, making this study a good candidate for a yoked control group procedure.

Participants in the control group were yoked in terms of how long the session lasted, so if a subject in the EMDR group took 25 minutes to reach a SUD level of 0–1, a subject in the yoked control group would participate in the control procedures for 25 minutes. The control group did everything the experimental group did (i.e., thought about the trauma, reported SUD), but instead of the eye movements, they focused their visual attention on a nonmoving red dot in the middle of a yellow card. By using the yoked control procedure, Dunn et al., (1996) guaranteed the average amount of time spent in a session would be identical for the experimental and control group subjects and that the two groups would do everything the same, except for the eye movements. They also began testing a third yoked group that would only think about the trauma but not get any form of therapeutic treatment during the session. After just a few subjects were tested, however, this third group was cancelled on ethical grounds—the subjects found the procedure too stressful.

The results? The EMDR group showed a significant reduction in the SUD score, as you can see in Figure 7.12.[6] Unfortunately for advocates of EMDR, the yoked control group also showed a drop and, while the reduction seems to be slightly larger for the EMDR group, differences between the two groups were not significant. That both groups showed essentially the same degree of improvement led Dunn et al. (1996) to conclude that a placebo effect is probably lurking behind any alleged success of EMDR.

One final point about this EMDR study is that it demonstrates that a finding of "no difference" between groups can be important. Recall from Chapter 4 that finding a significant difference

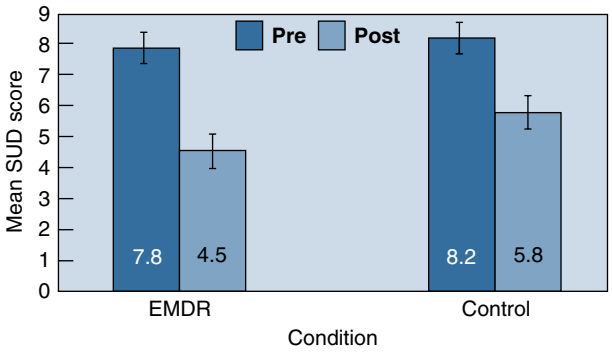

Figure 7.12
Results of the Dunn et al. (1996) study evaluating the effectiveness of EMDR therapy.

[6] You encountered *error bars* in Figures 5.1 in Chapter 5 and 6.2 in Chapter 6 and learned they have become standard features of graphs. They were not often used in the 1990s; Merikle and Skanes (1992) did not include them in their subliminal self-help graph (Figure 7.11), but Dunn et al., (1996) did in their EMDR graph (Figure 7.12).

when the null hypothesis is indeed false can be experimenter heaven. Failing to reject a null hypothesis is usually disappointing and can be a difficult finding to interpret, but such an outcome can be useful in a study like the EMDR one. Any time someone advocates a new treatment program, that person is obligated to show the program works. This means finding a significant difference between those getting the treatment and those not getting it, and a failure to find such a difference invites skepticism. Recall from Chapter 1 that researchers are skeptical optimists. They are prepared to accept new ideas supported by good research but are skeptical about claims not supported by empirical evidence.

SELF TEST

7.3

1. Look back at the hypothetical reaction time data in the "placebo control groups" section of the chapter. Suppose nothing but a placebo effect was operating. How would the reaction time numbers change?
2. In the Research Example evaluating EMDR, why is it that a finding of "no difference" between the experimental and control groups can be a useful outcome?

The designs in this chapter have in common the presence of a single independent variable. Some had two levels; some were multilevel. In Chapter 8, you will encounter the next logical step—designs with more than one independent variable. These are called *factorial designs*.

CHAPTER SUMMARY

Single Factor—Two Levels

The simplest experimental designs have a single independent variable (or factor) with two levels of that variable. These designs can include between-subjects factors or within-subjects factors. Between-subjects factors can be directly manipulated or they can be selected as subject factors. If manipulated, participants can be randomly assigned to groups (independent groups design) or matched on a potentially confounding variable, and then randomly assigned (matched groups design). If a subject variable is used, the between-subjects design is called an ex post facto design. Single-factor designs using a within-subjects factor are usually called repeated-measures designs (e.g., the famous Stroop studies). Studies using two levels of the independent variable are normally evaluated statistically with *t*-tests (assuming interval or ratio data, normal distributions, and homogeneity of variance).

Single Factor—More Than Two Levels

When only two levels of an experimental variable are compared, the results will always appear linear because a graph of the results will have only two points. Some relationships are nonlinear, however (e.g., the Yerkes-Dodson law; the Ebbinghaus forgetting curve), and they can be discovered by adding more than two levels to an independent variable. Adding levels can also function as a way to test and perhaps rule out (falsify) alternative explanations of the main result. Like the two-level case, multilevel designs can be either between- or within-subjects designs.

Analyzing Data from Single-Factor Designs

Results can be presented visually in a bar graph when the independent variable is a discrete variable, or in a line graph if the variable is continuous. Studies using more than two levels of an independent variable are normally evaluated statistically with a one-way analysis of variance or ANOVA (assuming interval or ratio data, normal distributions, and homogeneity of variance). A significant *F* ratio results in subsequent post hoc testing (e.g., Tukey's HSD test) to identify precisely which means differ. Independent groups and ex post facto designs are evaluated with a one-way ANOVA for independent groups; matched groups and repeated-measures designs are evaluated with a one-way ANOVA for repeated measures.

Special-Purpose Control Group Designs

In control group designs, the experimental treatment is absent for at least one condition. Varieties of control groups include placebo controls, often found in drug research; wait list controls, found in research on the effectiveness of a program or therapy; and yoked controls, in which the procedural experiences of the control group participants correspond exactly to those of the treatment group participants.

CHAPTER REVIEW QUESTIONS

1. Consider independent groups designs, matched groups designs, and ex post facto designs. What do they all have in common and how do they differ?
2. In the Research Example that examined a peer modeling program to help children with autism (Kroeger et al., 2007), why was a matched groups design used, instead of an independent groups design, and what was the matching variable?
3. Describe the Stroop effect, the experimental design used by Stroop (1935), and his method for controlling order effects.
4. Use the hypothetical caffeine and reaction time study to illustrate how multilevel designs can produce nonlinear effects.
5. Assuming for the moment the Yerkes-Dodson law is valid, explain why testing three levels of "arousal" yields a totally different result than testing just two levels.
6. Use the Bransford and Johnson (1972) experiment on the effects of context on memory to illustrate how a design with more than two levels of an independent variable can serve the purpose of falsification.
7. Describe when it is best to use a line graph and when to use a bar graph. Explain why a line graph would be inappropriate in a study comparing the reaction times of men and women, but was a good choice for the Ebbinghaus forgetting curve.
8. Describe the two varieties of two-sample *t*-tests and, with reference to the four designs in the first part of the chapter (single factor–two levels), explain when each type of test is used.
9. For an independent groups study with one independent variable and three levels, what is the proper inferential statistical analysis, and why is this approach better than doing multiple *t*-tests? How does post hoc testing come into play?
10. Use the example of the effects of alcohol on reaction time to explain the usefulness of a placebo control group.
11. Use the subliminal self-help study (Merikle & Skanes, 1992) to illustrate the usefulness of a wait list control group.
12. Use the study of the effectiveness of EMDR therapy (Dunn et al., 1996) to explain how a yoked control group works.

APPLICATIONS EXERCISES

Exercise 7.1. Identifying Variables

As a review, look back at each of the Research Examples in this chapter and identify: (a) the independent variable; (b) the levels of the independent variable; (c) the type of independent variable (situational, instructional, task, and subject); (d) the dependent variable; and (e) the scale of measurement of the dependent variable (nominal, ordinal, interval, or ratio).

1. Research Example 11—Two-Level Independent Groups Design
2. Research Example 12—Two-Level Matched Groups Design
3. Research Example 13—Two-Level Ex Post Facto Design
4. Research Example 14—Two-Level Repeated Measures Design
5. Research Example 15—Multilevel Independent Groups Design
6. Research Example 16—Multilevel Repeated Measures Design
7. Research Example 17—Using Both Placebo and Wait List Control Groups
8. Research Example 18—A Yoked Control Group

Exercise 7.2. Identifying Designs

For each of the following descriptions of studies, identify the independent and dependent variables involved and the nature of the independent variable (between-subjects or within-subjects; manipulated or subject variable), name the experimental design

being used, identify the measurement scale (nominal, ordinal, interval, or ratio) for the dependent variable(s), and indicate which inferential analysis ought to be done (which type of *t*-test or ANOVA).

1. In a study of how bulimia affects the perception of body size, a group of woman with bulimia and a group of same-age women without bulimia are asked to examine a precisely graded series of 10 drawings of women of different sizes and to indicate which size best matches the way they think they look.

2. College students in a cognitive mapping study are asked to use a direction finder to point accurately to three unseen locations that differ in distance from the laboratory. One is a nearby campus location, one is a nearby city, and the other one is a distant city.

3. Three groups of preschoolers (50 per group, assigned randomly) are in a study of task perseverance in which the size of the delay of reward is varied. The children in all three groups are given a difficult puzzle and told to work on it as long as they would like. One group is told that as payment they will be given $5 at the end of the session. The second group will get the $5 after two days from the end of the session, and the third will get the money after four days.

4. To examine whether crowding affects problem-solving performance, participants are placed in either a large or a small room while attempting to solve a set of word puzzles. Before assigning participants to the two conditions, the researcher takes a measure of their verbal intelligence to ensure the average verbal IQ of the groups is equivalent.

5. In a study of first impressions, students examine three consecutive photos of a young woman whose arms are covered with varying amounts of tattoos. In one photo, the woman has no tattoos; in the second photo, she has one tattoo on each arm; in the third photo, she has three tattoos per arm. From a checklist, students indicate which of five majors the woman is likely to be enrolled in and rate her on 10 different 7-point scales (e.g., one scale has 1 = emotionally insecure and 7 = emotionally secure).

6. In an attempt to identify the personality characteristics of cell phone users, three groups of college students are identified: those who do not have a cell phone; those who own a cell phone, but report using it less than 10 hours per week; and those who own a cell phone and report using it more than 10 hours per week. They are given a personality test that identifies whether they have an outgoing or a shy personality.

7. A researcher studies a group of 20 men, each with the same type of brain injury. They are divided into two groups in such a way that their ages and educational levels are kept constant. All are given anagram problems to solve; first group is given 2 minutes to solve each anagram and the second group is given 4 minutes per anagram.

8. To determine if maze learning is affected by the type of maze used, 20 rats are randomly assigned to learn a standard alley maze (i.e., includes side walls; located on the lab floor); another 20 learn an elevated maze (no side walls; raised above floor level). Learning is assumed to occur when the rats run through the maze without making any wrong turns.

Exercise 7.3. Outcomes

For each of the following studies, decide whether to illustrate the described outcomes with a line graph or a bar graph; then create graphs that accurately portray the outcomes.

1. In a study of the effects of marijuana on immediate memory for a 30-item word list, participants are randomly assigned to an experimental group, a placebo control group, or a straight control group.

   Outcome A. Marijuana impairs recall, while expectations about marijuana have no effect on recall.

   Outcome B. Marijuana impairs recall, but expectations about marijuana also reduce recall performance.

   Outcome C. The apparently adverse affect of marijuana on recall can be attributed entirely to placebo effects.

2. A researcher uses a reliable and valid test to assess the autonomy levels of three groups of first-year female college students after they have been in college for two months. Someone with a high level of autonomy has the ability to function well without help from others—that is, to be independent. Tests scores range from 0 to 50, with higher scores indicating greater autonomy. One group (R300) is made up of resident students whose homes are 300 miles or more from campus; the second group includes resident students whose homes are less than 100 miles from campus (R100); the third group includes commuter students (C).

   Outcome A. Commuter students are more autonomous than resident students.

   Outcome B. The farther one's home is from the campus, the more autonomous that person is likely to be.

   Outcome C. Commuters and R300 students are both autonomous, while R100 students are not.

3. Animals learn a maze and, as they do, errors (i.e., wrong turns) are recorded. When they reach the goal box on each trial, they are rewarded with food. For one group of rats, the food is delivered immediately after they reach the goal (0 delay). For a second group, the food appears 5 seconds after they reach the goal (5-second delay).

Outcome A. Reinforcement delay hinders learning.

Outcome B. Reinforcement delay has no effect on learning.

4. Basketball players shoot three sets of 20 foul shots under three levels of arousal: low, moderate, and high. Under low arousal, every missed free throw means they have to run a lap around the court (i.e., it is a minimal penalty, not likely to cause arousal). Moderate arousal means two laps per miss and high arousal means four laps per miss (i.e., enough of a penalty to create high arousal, perhaps in the form of anxiety). It is a repeated-measures design; assume proper counterbalancing.

Outcome A. There is a linear relationship between arousal and performance; as arousal increases, performance declines.

Outcome B. There is a nonlinear relationship between arousal and performance; performance is good only for moderate arousal.

Answers to Self Tests

✓7.1

1. Single factor, two-level independent groups design.
2. Single factor, two-level repeated-measures design.
3. An ex post facto design.

✓7.2

1. Independent samples *t*-test.
2. The IV is a discrete variable, with no intermediate points that would allow for extrapolation in a line graph.
3. Repeated measures one-way ANOVA, with a post-hoc test if F-test is statistically significant.

✓7.3

1. Instead of 0.32 sec, the RT for the experimental group would be 0.22 sec (same as placebo group).
2. It raises questions about the validity of the new therapy being proposed. Those making claims about therapy are obligated to show that it works (i.e., produces results significantly better than a control group).