CHAPTER 7

Experimental Design I: Single-Factor Designs

Preview & Chapter Objectives

Chapters 5 and 6 have set the stage for this and the following chapter. In Chapter 5, I introduced you to the experimental method; distinguished 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 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 various experimental designs that yield the productions 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 subject of Chapter 8. When you finish this chapter, you should be able to:
Identify the four varieties of single-factor designs: independent groups, matched groups, nonequivalent groups, repeated measures.

Know when to use a t test for independent groups and when to use a t test for dependent groups.

Understand the logic behind the use of the three special types of control groups: placebo, waiting list, and yoked.

Understand the ethical issues involved with the use of control groups.

Explain the two different 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.

Understand why multiple t tests are inappropriate and a 1-way ANOVA is appropriate when analyzing data from single-factor, multilevel studies.

In Chapter 3's discussion of scientific creativity (p. 94), I used the origins of maze learning research as an example. Small's research, using a modified version of the Hampton Court maze, was just 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 Chapter 2 (Box 2.3, p. 56) 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 or kinesthetic sense to learn and recall the maze. In effect, the rat learns to take so many steps, then turn right, and so on. To test his kinesthesia idea directly, he completed a simple yet elegant study with his University of Chicago colleague Harvey Carr (Carr & Watson, 1908). After one group of rats learned a complicated maze, Carr and Watson removed a middle section of the maze structure, thereby making certain portions of the maze shorter than before. They predicted that rats trained on the longer maze might literally run into the wall when the maze was shortened. Sure enough, in a description of one of the rats, Carr and Watson 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 side wall 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 kinesthesia hypothesis, but the important point here is that good research does not require immensely complex research designs. In some cases, two groups will do just fine.

**Single Factor—Two Levels**

As you can see from the decision tree in Figure 7.1, there are four basic designs involving one independent variable that has two levels, and they result from a series of decisions about the independent variable being investigated. First, this variable
can be tested 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 matching followed by random assignment is needed. As you recall from Chapter 6, decisions about matching have to do with sample size and the need to be especially careful about extraneous variables that are highly correlated with the dependent variable. If a subject variable is being investigated, the groups are composed of different categories of individuals (e.g., male/female, introverted/extroverted, liberal/conservative). The design is sometimes called an “ex post facto” design because the groups are formed “after the fact” of their already-existing subject characteristics; it has also been referred to as a “natural groups” design and as a nonequivalent groups design, the term I will use. Because groups are made up of different types of individuals in this kind of design, researchers using them often try to reduce the nonequivalence as much as possible by matching the groups on a variety of factors. For instance, a nonequivalent groups study comparing males and females might ensure that the participants in each group are about the same age and from the same socioeconomic class.

The final 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 some specific examples.
**TABLE 7.1** Attributes of Four Single-Factor Designs

<table>
<thead>
<tr>
<th>Type of Design</th>
<th>Minimum Levels of Independent Variable?</th>
<th>Independent Variable Between or Within?</th>
<th>Independent Variable Type?</th>
<th>Creating Equivalent Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independent groups</td>
<td>2</td>
<td>between</td>
<td>manipulated</td>
<td>random assignment</td>
</tr>
<tr>
<td>Matched groups</td>
<td>2</td>
<td>between</td>
<td>manipulated</td>
<td>matching</td>
</tr>
<tr>
<td>Nonequivalent groups</td>
<td>2</td>
<td>between</td>
<td>subject</td>
<td>matching may reduce nonequivalence</td>
</tr>
<tr>
<td>Repeated measures</td>
<td>2</td>
<td>within</td>
<td>manipulated</td>
<td>n/a</td>
</tr>
</tbody>
</table>

**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 usually produce more elaborate and more intriguing outcomes. Also, few journal editors are impressed 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 are case examples of three such experiments.

**Case Study 6—Independent Groups**

An example of an independent groups design using a single factor with two levels is a well-known study by Blakemore and Cooper (1970). They were interested in the effects of experience on the development of the visual system. Two-week-old cats were randomly assigned to the two levels of a manipulated independent variable that could be called “visual environment.” The cats were to be raised in a setting dominated by either horizontal or vertical stripes. I think you can see why this had to be a between-subjects rather than a within-subjects design; it wouldn’t make sense to raise a kitten in a vertical environment and then in a horizontal environment. In studies like this one, participants experiencing one level of the independent variable are in essence “used up”; the experience makes it impossible for them to “start over” in the experiment’s other condition.

Figure 7.2 shows Blakemore and Cooper’s sketch of the apparatus for the “vertical” condition. The cat is standing on Plexiglas, and the stripes extend above and below the surface. The wide collar around the cat’s neck keeps the animal visually focused on the walls of the chamber. Over a period of several months, the cats were exposed to either the vertical or the horizontal world for 5 hours per day; they were kept in a darkened environment otherwise.

At the end of the study, Blakemore and Cooper tested the animals both behaviorally and by measuring the activity of neurons in their visual cortex. In general, the cats recovered quickly from the deprivation. After “10 hours of normal vision they ... would jump with ease from a chair to the floor” (Blakemore & Cooper, 1970, p. 477). However, the cats raised in a vertical environment apparently could...
Creating Equivalent Groups

You might think, produce more are impressed beauty in sim-
with two levels is tered in the interested in the
for the “verti-
for normal vision
universal vision.

not perceive horizontal events very well; likewise, horizontally raised cats had prob-
problems with vertical stimuli:
The differences were most marked when two kittens, one horizontally and
the other vertically experienced, were tested simultaneously with a long black
or white rod. If this was held vertically and shaken, the one cat would follow
it and play with it. Now if it was held horizontally the other cat was attracted
and its fellow ignored it (Blakemore & Cooper, 1970, p. 478)
Obviously, early experience can profoundly affect how the brain develops.

Case Study 7—Matched Groups
Old movies sometimes feature attempts to extract information from a hero who has
been locked up, perhaps tortured, and deprived of sleep for 2 or 3 days. Can sleep
deprivation affect the responses to interrogation questions? That was the empirical
question asked by Blagrove (1996) in an interesting study using a matched groups
design. More specifically, he wanted to know if sleep-deprived people would be
influenced by misleading questions. He recruited college students for three separate
studies, each involving two groups—some participants were sleep deprived and oth-
ers weren’t. The sleep-deprived students spent their time in the laboratory, kept
awake for 21 straight hours in the first two studies and 43 hours in the third. Close
monitoring by “20 shifts of research assistants” (p. 50) ensured that they remained
awake. Nondeprived students were allowed to sleep at home. The matching variable
was “self-reported habitual sleep durations” (p. 50). Blagrove wanted to make sure
that the typical length of sleep was held constant for the two groups to “control for
sleep length related group differences in personality and sleep-stage characteristics”
(p 50) So the average self-reported sleep times for the two groups were 8.4 and 8.5 hours for the first study, 8.3 and 8.1 for the second, and 8.4 and 8.1 for the third.

All participants in the study were given a standardized suggestibility test in which they listened to a story and then responded to leading questions about it (i.e., questions that could not be directly answered from information given in the story). After responding, they were given negative feedback about their responses and then questioned again to see if they changed any answers. In general, they were more influenced by the leading questions and more likely to change their answers after being sleep deprived, especially in the third study, in which the sleep deprivation lasted for 48 hours. And using the matching procedure, to ensure that the groups were similar in their typical sleeping patterns, made it possible to attribute the group differences to sleep deprivation.

Case Study 8—Nonequivalent Groups
Stimulated perhaps by Terman’s mega-longitudinal study of gifted children (see Box 6.1, pp. 200–201), considerable research over the years has attempted to shed light on the gifted child. A study by Knepper, Obrutz, and Copeland (1983) asked whether gifted children might be adept at social and emotional problem solving, in addition to their normal advantages over average children in cognitive problem solving. Their experiment nicely illustrates a nonequivalent groups design. The independent variable was a subject variable, degree of giftedness, and two levels were compared, gifted (operationally defined as IQ = 130 or higher) and average (IQ between 90 and 110) students. The mean IQs for the groups were 136.9 and 102.9, respectively. A specific matching procedure was not used, but age was controlled by using only sixth graders. All were given a test called the Means-Ends Problem Solving Test, which measures the quality of solutions to interpersonal (social) and intrapersonal (emotional) problems. The gifted children indeed outperformed the average children on these tests of social and emotional problem solving, a finding consistent with Terman’s conclusion that gifted children are not just “brains,” but also have some social skills.

One important caution. Recall from Chapter 5 (pp. 155–156) that conclusions about cause and effect cannot be drawn when subject variables are involved. Thus, it would be inappropriate to say that giftedness somehow caused an increase in the ability to solve and emotional problems. All that can be said is that the gifted and nongifted children differed in how well they solved such tasks.

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 sequence problems. 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 then B, and the rest will get B then A. If participants are tested more than once per condition, reverse counterbalancing (ABBA) can 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 (and to learn more about swastikas than you probably know at the moment), read Box 7.1 before continuing.

Another counterbalancing strategy for a study with just two conditions, when each condition is being tested many times, is simply to alternate the conditions (ABAB...). Such an approach was taken in the following case study.

---

**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 first 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. What did Stroop do?

The original study summarized three experiments, completed by Stroop as his doctoral dissertation. We’ll focus on the first two 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 36 females performed two tasks. Both involved reading the names of colors. Stroop (1992, p. 16) called one of the conditions RCNb (“Reading Color Names printed in black”). 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 (“Reading Color Names where the color of the print and the word are different”). 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 was aware of the problems with sequence 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.

Stroop’s experiment 1 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, then, 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 (“Naming Color test”), participants named the colors of square color patches. In the second and key condition, NCWd (“Naming Color of Word test where the color of the print and the word are different”), 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.” Participants in 1935 had the same difficulty experienced by people 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 that the average color naming times were 63.3 seconds for condition NC and a whopping 110.3 seconds for the NCWd condition. I’ve taken the four different outcomes, reported by Stroop in the form of tables, and drawn a bar graph of them in Figure 7.3. As you can see, the Stroop effect is a robust phenomenon.

I 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 interesting aspect of this final study was that in the place of square color patches on the NC test, Stroop substituted color patches in the shape of swastikas, which “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 swastika is an ancient religious symbol formed by bending the arms of a traditional Greek cross (+). Ironically, Stroop’s study was published the same year (1935) that the swastika became officially adopted as the symbol for Nazi Germany.

![Figure 7.3 Combined data from the first two experiments of the original Stroop study (1935)]
Case Study 9—Repeated Measures

In a study of motion perception and balance, Lee and Aronson (1974) tested some predictions from a theory of perception proposed by James Gibson, mentioned briefly in Chapter 1 as the husband of Eleanor Gibson. In particular, they were interested in how we maintain our balance in a moving environment. Children aged 13 to 16 months were placed in the apparatus pictured in Figure 7.4. When an infant was facing the back wall, the experimenter could move the walls and ceiling either forward or backward.

It was hypothesized that moving the room forward (Figure 7.5a) would create an “optic flow pattern” identical to the one produced if the infant’s head was moving backward (Figure 7.5b). This, in turn, would trigger a compensating tilt forward by the child. If so, then moving the room forward should cause the infant to lean forward or perhaps fall forward (Figure 7.5c). Just the opposite was predicted when the room was moved backward.
Unlike the study with the cats raised either in a vertical or a horizontal world, there is no reason why Lee and Aronson’s infants could not experience both experimental conditions: the room moving forward or backward. Hence, a within-subjects approach was taken: the design was a single-factor repeated-measures design. The independent variable was the direction of the room’s movement, either forward or backward, and the infants’ body lean or falling was measured as the dependent variable. Twenty repeated trials were completed per subject, with the room’s movement alternating from trial to trial. For some children, the alternating sequences began with the room moving forward; for others, the room moved backward on their first trial. Seven participants were tested, ranging in age from 13 to 16 months, but three of them became distressed and for them the experiment was terminated. The responses of the remaining four participants were recorded by three observers (why would more than one observer be needed?) Some loss of balance in the predicted direction occurred on 82% of the trials. Loss of balance was categorized by the observers as a sway (26% of the trials), a stagger (23%), or a fall (33%—ouch!).

One drawback to a counterbalancing procedure that simply alternates between conditions A and B is that subjects can easily predict what condition is about to occur. However, Lee and Aronson correctly decided that this problem was unlikely to influence their results, given the ages of their participants. Another reason for their choice of counterbalancing was practical: with the infants remaining in the moving room throughout the experimental session, once the room had been moved in one direction, the next trial had to be a movement in the opposite direction.

Analyzing Single-Factor, Two-Level Designs

To determine whether the differences found between the two conditions of a two-level design are significant or due simply to chance, some form of inferential statistical analysis is required. When interval or ratio scales of measurement are used in the experiment, the most common approach is to use one of two varieties of the t test, a procedure mentioned near the end of Chapter 4 and described in more detail in Appendix C. Other techniques are required when nominal or ordinal scales of measurement are used.

There are two forms of the t test. The first is called a test for independent groups, and as the name implies, it is used when the two groups of participants are completely independent of each other. This occurs (a) whenever participants in the
study are randomly assigned to the groups or (b) if the variable being studied is a subject variable (e.g., males vs. females). If the independent variable is a within-subjects factor, or if two separate 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 *t test for dependent groups* (sometimes called a *t test for correlated groups*) is used. For the four single-factor designs just considered, the following *t* tests would be appropriate:

- *t* test for independent groups
  - independent groups design
  - nonequivalent groups design
- *t* test for dependent groups
  - matched groups design
  - repeated-measures design

In essence, the *t* test examines the difference between the two mean scores and determines (with some probability) whether this difference is larger than expected by chance factors alone. If it is larger, and if potential confounds can be ruled out, then the researcher can conclude with a high probability that the differences are real, get the study published, and perhaps get promoted. See Appendix C for step-by-step instructions on how to carry out both varieties of *t* tests, as well as an accompanying analysis of effect size.

**Control Group Designs**

I introduced the basic distinction between experimental groups and control groups in Chapter 5 (p. 150). Experimental groups receive some treatment, while those in the control group do not. To expand the logic to repeated-measures designs, which do not have different groups, a parallel distinction can be made between experimental conditions and control conditions. Besides the typical control group situation in which a group is left untreated, there are three other specific types of control groups worth describing: placebo controls, waiting list controls, and yoked controls.

**Placebo Control Groups**

A "placebo" is a substance that appears to have a specific effect but in fact is pharmacologically inactive. Sometimes patients will feel better when given a placebo but told it is drug X, 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 some treatment when in fact they aren't. You can see why this would be necessary. Suppose you wished to determine if alcohol slows down 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 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 subtly influenced by that knowledge. To solve this biasing problem, you should 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—these individuals yield a simple baseline measure of reaction time. If you use all three groups and get these average reaction times:

- experimental group: 32 seconds
- placebo control: 22 seconds
- straight control: 16 seconds

you could conclude that what people expect about the effects of alcohol slows reaction time somewhat (from .16 to .22) but that alcohol by itself also has an effect beyond people's expectations (.22 to .32).

An example of a study with both a placebo control group and a straight control group examined the effects of parasitic infection on various cognitive and motor skills in Jamaican children (Sternberg, Powell, McGrane, & Grantham-McGregor, 1997). Fourth and fifth graders with mild infections were randomly assigned (after matching for gender) to a drug treatment group and a placebo group. A third (non-equivalent) group of children did not have infections. The drug effectively eliminated the infection and was administered to the placebo children immediately after the end of the study, but the infection's consequences lingered. Compared to the untreated, uninfected controls, children in both the drug group and the placebo group did poorly on several cognitive tasks. Hence, while the drug might have an important medical effect, it did little to improve the cognitive deficits that accompany the illness. Because the children lived in an environment where parasitic infections are commonplace, Sternberg et al. recommended that a program of cognitive remediation be added to the medical intervention.

### Waiting List Control Groups

**Waiting list control groups** are often used in research designed to assess the effectiveness of some program (Chapter 10) 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 some type of problem that the program is supposed to alleviate. 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 that the clients in all three groups were generally equivalent to each other, the control group subjects also had to be nightmare sufferers. Those assigned to the waiting list were assured that they would be helped, and after the study ended they were given treatment equivalent to that experienced by the experimental groups.

Giving the waiting list participants an opportunity to benefit from some therapy procedure provides an important protection for participant welfare, but it also creates pressures on the researcher to use this control procedure only for therapies or programs of relatively brief duration. In fact, some might argue that it is unethical to put people into a waiting list control group because they won't receive the program's benefits right away. This issue can be especially problematic when research
problem, you cannot be dis- is the placebo drinks at all)? action time. If

ol slows reac- has an effect
right control ve and motor m-McGregor, assigned (after A third (non-actively elimi- neated immediately after pared to the d the placebo might have an s that accom- parasitic infec- n of cognitive

Box 7.2

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 just told to learn the word list, 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 on the effects of a 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 that one of your relatives had been assigned to the control group, do you think you would be concerned?

In a similar vein, there has been some 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., they are in the control group). Some researchers argued that 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 7.6? 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 after the fact to say 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 would be effective. The only way to tell is to do the study. Prior to Langer and Rodin’s nursing home study, for example, one easily could have predicted that the experimental 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 likely that control group participants really felt left out or deprived.

Second, researchers point out that in research evaluating a new treatment or program, the comparison is seldom between the new treatment and no treatment; it is usually between the new treatment and the current treatment. So for control group members, available services are not really being withheld; they are receiving the normal services. Furthermore, once the study has demonstrated some positive effect of
Figure 7.6  Potential consequences of being assigned to the control group.

In the experimental treatment, members of the control groups are typically given the opportunity to be treated.

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, then, programs with empirically demonstrated effectiveness serve the general good and may in many cases save or prolong lives.

Evaluates life-influencing programs. Read Box 7.2 for an examination of this issue and a defense of the use of control groups in research.

Remember the Chapter 1 discussion of pseudoscience, as illustrated with the case of subliminal self-help audiotapes? As we saw, 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 case study, which effectively combines both placebos and waiting lists to yield a new interpretation of what these tapes accomplish.

Case Study 10—Using Both Placebo and Waiting List Control Groups
One of the favorite markets for the subliminal self-help audiotape business is in the area of weight loss. Americans in particular try to lose weight by trying 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 subliminal tapes—just pop in a tape 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 waiting list control, Merckle and Skanes (1992) evaluated the effectiveness of these self-help weight loss tapes. 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 commercial subliminal self-help tape 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 were given one designed to relieve dental anxiety (the researchers had a sense of humor). The two tapes were indistinguishable to ordinary listeners. A third group, the waiting 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 1 to 3 hours per day and participants in all three groups were weighed weekly for five weeks. The results? Those in the experimental group lost a modest amount of weight, but the same amount was also lost by the placebo control group. This is the typical outcome with this type of study, indicating that the subliminal tapes have no effect by themselves. The interesting outcome, however, was that the waiting list group also lost weight, about the same amount as the other two groups. This led Merckle 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 minds were being altered, would have lost weight, but the waiting list group folks, not yet in possession of the tapes, would not have lost any weight. Their outcome led the authors to argue that the effect of the tapes was to focus a person’s attention on the problem at hand, in this case weight loss. The subjects in all 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, then, the waiting list 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 it, the study’s outcome also sounds suspiciously like a Hawthorne effect, which you learned about in Chapter 6 (p. 204).

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 Merckle, arranged the 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). That is, the authors built a nice double-blind control into their study.

**Yoked Control Groups**

A third type of control group is the yoked control group. It is used when participants in the experimental group, for one reason or another, participate for varying amounts of time or are subjected to different types of events in the study. Each member of the control group is matched or “yoked” to a member of the experimental
Fig 7.7 The experimental setup of the Weiss (1968) study, illustrating the use of a yoked control group.

Case Study 11—Yoked Controls and Stress

A good example of a yoked control group is a study by Weiss (1968) on the relationship between control of stress and health. Experimental rats were exposed to occasional mild shocks to their tails, which they could turn off (i.e., control) or avoid altogether by rotating a small wheel with their paws (Figure 7.7). A control group of rats was never shocked. Rats in the middle position of the apparatus were in the yoked control group. Each rat in this group was yoked to (paired with) a rat in the experimental group so that it received the exact same number of shocks but could not control them. On a given trial, a tone would sound, which could be heard by all three animals. The rat in the experimental group then had 10 seconds to turn the wheel. If it failed to do so, it received a mild shock to its tail, and the yoked rat also received a
shock (regardless of what it happened to be doing at the time). Thus, the rats in the experimental and yoked control groups received exactly the same degree of aversiveness (i.e., same number of shocks), but they differed in whether they had control over the situation. Weiss concluded that having control over the shock helped the rats avoid some of the unhealthy consequences of stress. The yoked rats were more likely to develop ulcers and to lose weight, while rats in the experimental group and the unshocked rats were relatively unaffected by their experience.

You might recognize the Weiss design. It was also used by Seligman’s research team in their studies on learned helplessness. If you look back to the discussion of theory in Chapter 3 (pp. 82–84), you will find a description of a yoked control procedure in Seligman’s triadic design, although I didn’t use the term “yoked” at that point (except in a footnote). Another example of a yoked control was the Brady study of ulcers in executive monkeys, the focus of Box 5.3 (pp. 170–171). In fact, Weiss’s study was designed in part to correct for the methodological deficiencies (i.e., subject selection problems) in Brady’s research (Weiss, 1977).

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; for that reason, they often are called single-factor multilevel designs. As was true for two-level designs, these include both between- and within-subjects designs and can be of the same four types: independent groups, matched groups, nonequivalent groups, and repeated measures.

Between-Subjects, Multilevel Designs

A distinct advantage of multilevel designs is that they enable the researcher to discover nonlinear effects. To take a simple between-subjects example, suppose you were interested in the effects of caffeine dosage on reaction time. You set up an experiment that compares two dosage levels (1 mg and 3 mg), get the results in Figure 7.8, and conclude that as a stimulant, caffeine speeds up reaction time. As the dosage increases, reaction time quickens in a straight-line (i.e., linear) fashion.

Now suppose another researcher does this study but uses a multilevel design that includes four dosages (1 mg, 2 mg, 3 mg, and 4 mg)—an example of replicating (1 mg and 3 mg) and extending (2 mg and 4 mg) your finding. The study might produce the results in Figure 7.9.

This outcome replicates your results for the 1-mg and 3-mg conditions exactly, but the overall pattern for the four conditions calls your conclusion into serious question. Instead of caffeine simply reducing reaction time, the conclusion now would be that (a) adding caffeine quickens reaction time, but only after a level of 2 mg is reached and (b) caffeine reduces reaction time only up to a point; after 3 mg, caffeine begins to slow down reaction time. That is, the outcome is no longer a simple linear result but rather a nonlinear one. In general, then, the advantage of multi-
level designs is that they are more informative and often provide for more interesting outcomes than two-level designs.

One of psychology's most famous graphs illustrates a nonlinear effect. It shows how the passage of time influences the rate of forgetting and yields some insight into the common feeling one gets two days after an exam—"I don't remember anything!" The curve comes from the pioneering memory research of Hermann Ebbinghaus, as described briefly in Box 7.3.

In addition to identifying nonlinear relationships, single-factor multilevel designs can also test for specific alternative hypotheses and perhaps rule them out. This is a strategy you will recognize from the discussion in Chapter 3 (pp. 87–88) of the merits of falsification. A perfect example is a study by Bransford and Johnson (1972).

Case Study 12—Multilevel Independent Groups
Cognitive psychologists interested in how we comprehend new information have shown that understanding new ideas is easier if they are placed in some context. For example, understanding a textbook chapter is easier if you first read a preview and set of objectives, as I hope you have already discovered as you have been using this text. The Bransford and Johnson study illustrates these context effects. In their study, participants were asked to comprehend this paragraph. Try it yourself:

If the balloons popped, the sound wouldn't be able to carry, since everything would be too far away from the correct floor. A closed window would also
Box 7.3

**ORIGINS—Nonlinear Results: The Ebbinghaus Forgetting Curve**

The 19th-century German psychologist Hermann Ebbinghaus (1850–1909) is justifiably famous for his pioneering research on memory and forgetting. At a time when psychology was in its infancy and “how to do psychological research” guides did not exist, Ebbinghaus managed to complete an extended series of studies that set a standard for precision and methodological rigor. His purpose was to examine the formation of associations in the mind, and his first task was to find materials that were devoid of associations. His solution, considered one of the best examples of creativity in scientific psychology, was to string together sequences of consonants, vowels, and consonants. These CVCs are more widely known as nonsense syllables; Ebbinghaus created about 2,300 of them. For several years, showing either tremendous perseverance or a complete lack of any social life, Ebbinghaus spent hours at a time memorizing and trying to recall various lists of these CVCs. Yes, he was the only participant. He systematically varied such factors as the number of syllables per list, the number of study trials per list, and whether the study trials were crammed together or spaced out. He published his results in a small monograph called *Memory: A Contribution to Experimental Psychology* (1885/1964).

In his most famous study, the one resulting in a nonlinear outcome, Ebbinghaus examined the time course of forgetting. His empirical question was this: Once some material has been memorized, how much of that memory persists after varying amounts of time? His procedure was to first commit to memory eight 13-item lists of nonsense syllables, then wait for some period of time, then attempt to relearn the lists. His time intervals were 20 minutes, 1 hour, 9 hours, 1 day, 2 days, 6 days, and 31 days. Ebbinghaus recorded the total time for both the original learning of the eight lists and the relearning of the lists. Original learning minus relearning yielded a measure of “saving,” which was converted to a percentage by dividing by the time of original learning. Thus, if the original learning took 20 minutes and relearning took 5 minutes, 15 minutes or 75% (15/20 x 100) of the original learning time was saved.

His results, which you can probably find in the memory chapter of your general psychology text, are shown in Figure 7.10. As you can see, recall declined with the passage of time, but the decline was not a steady or linear one. Clearly, it is a nonlinear effect. Forgetting occurs very rapidly at first, but then the rate of forgetting slows down. 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%).

From the standpoint of methodological control, there are several other interesting things about the Ebbinghaus research. To ensure a constant presentation rate, for example, Ebbinghaus set a metronome to 150 beats per minute and read each CVC...
exactly 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 pure repetition. Also, he worked only when sufficiently motivated so that he could “keep the attention concentrated on the tiresome task and its purpose” (Ebbinghaus, 1885/1964, p. 25).

![Graph showing the Ebbinghaus forgetting curve](image)

**Figure 7.10** The Ebbinghaus forgetting curve—a non-linear outcome

prevent the sound from carrying, since most buildings tend to be well insulated. Since the whole operation depends on a steady flow of electricity, a break in the middle of the wire would also cause problems. Of course, the fellow could shout, but the human voice is not loud enough to carry that far. An additional problem is that a string could break on the instrument. Then there could be no accompaniment to the message. It is clear that the best situation would involve less distance. Then there would be fewer potential problems. With face to face contact, the least number of things could go wrong. (Bransford & Johnson, 1972, p. 392)

I imagine your reaction to this passage is “Huh?”—a response shared by many of the participants in the original study. However, Bransford and Johnson found that comprehension could be improved by adding some context. Here’s what they did.

They designed a single-factor independent groups study with five levels of the independent variable. Participants randomly assigned to a control group were asked to do what I just asked of you. They read the paragraph and tried to recall as much as they could of the 14 idea units included in it. They recalled an uninspiring average of 3.6 ideas. A second group read the story twice to see if recall might improve with simple repetition. It didn’t—they recalled 3.8 ideas. A third group was first given a look at the cartoon in Figure 7.11a. Then they read and tried to recall the paragraph. They recalled 8.0 ideas out of 14. Clearly, the cartoon gave these participants an overall context within which to comprehend the sentences of the para-
be well insu-

duced by many of

is not true that

on found that

five levels of the

group were asked

the same amount

gave these partici-

 helped improve and group was first

tried to recall the
gave these partici-

graph. But is it necessary to see the cartoon first, before reading the passage? Yes. The
fourth condition of the study had participants read the paragraph, then see the car-
toon, then recall the paragraph. They recalled 3.6 ideas, just like the control group.
Finally, the fifth group was given a partial context. Before reading the passage, they
saw the cartoon in Figure 7.11b. It contains all the elements of 7.11a, but
rearranged. Participants in this group recalled an average of 4.0 ideas. In graph form,
the results looked like Figure 7.12.

Considering just two groups, “No Context-1 Repetition” and “Context Before,”
this study is reasonably interesting, showing a simple improvement in compre-
ッション by adding some context in the form of the cartoon. However, adding the other
conditions makes it a really interesting study by ruling out (i.e., falsifying) some
alternative factors that might be thought to improve recall. Thus, context improves
our understanding of something but only if that context occurs first. Because pre-
senting the context after the reading doesn’t help, it can be inferred that context improves recall by facilitating the initial processing of the information, not just its subsequent retrieval. Also, the argument that simple repetition would improve performance can be ruled out—doubling the repetitions didn’t improve performance. Lastly, it isn’t enough just to present all the individual elements of the story (as in the partial context condition)—the elements have to be arranged in a way that relates meaningfully to the material to be recalled.

**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 just 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 block randomization procedures can be used. In the following case study, each condition was tested just once and a Latin square was used to accomplish counterbalancing.

**Case Study 13—Multilevel Repeated Measures**

Can listening to Mozart make you smarter? Some people apparently think so, and the phenomenon has been dubbed the “Mozart Effect.” Despite the lack of any consistent evidence to support the idea, it has been promoted to parents as a way to give their children an edge in the IQ wars. As you might guess, there’s even a Mozart Effect website (www.mozarteffect.com) that features a variety of tapes and books for sale. In true pseudoscience fashion, the site argues for the phenomenon’s effectiveness through the use of testimonials and anecdotal evidence. In a description of one of the tapes for sale on the website, which features Mozart’s violin concertos, the website claims that listening to the concertos will “increase verbal, emotional and spatial intelligence, improve concentration and memory, and strengthen intuitive thinking skills,” and that the “high frequencies of the violins exercise the ears so the stimulation to the brain is balanced by the wonderful harmonies” (whatever that means). So what is the basis for these extraordinary claims?
The origin appears to be a brief study published in 1993 in the journal *Nature*, in which the researchers evidently showed that listening to 10 minutes of Mozart could produce a short-term (i.e., the effect didn't last) increase in spatial reasoning ability among college students (Rauscher, Shaw, & Key, 1993). Cognitive psychologists were skeptical of even this short-term effect and numerous unsuccessful attempts were made to replicate the finding. One such study was completed by Steele, Ball, and Runk (1997).

Steele and his colleagues included three conditions in their study: listening to Mozart for 10 minutes, listening to a recording of soothing environmental sounds (e.g., a gentle rainstorm) for 10 minutes, and not listening to anything—sitting quietly for 10 minutes and trying to relax. All 36 participants were tested in each of the three conditions, making this a within-subjects, multilevel design. Although complete counterbalancing would have been easy to implement (six different sequences of conditions, six participants randomly assigned to each sequence), the authors chose to use a $3 \times 3$ 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, they were "told that the experiment concerned the effect of relaxation on recall" (Steele et al., 1997, p. 1181). The memory task was a reversed digit span procedure: 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, then be given three consecutive digit span tasks. Each digit span included 9 numbers, presented in a random order. Thus, a score from 0 to 27 could be earned.

The study did produce some statistically significant findings, but none that would comfort those marketing the Mozart effect tapes. The average number of digits correctly recalled was virtually identical for all three conditions: 18.53 for the Mozart tape, 18.50 for the gentle rain tape, and 18.72 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. So, should people play Mozart tapes for their children? Of course—it's wonderful music. Will it make them smarter? Apparently not, although it could make them enjoy classical music, an outcome of value by itself.

### 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 just two or three levels (e.g., the Mozart example) but makes for tedious reading as the amount of data increases. You might have noticed this when reading about the results of the five conditions of the Bransford and Johnson (1972) study. A second approach is to construct a table of results. A table for the Bransford and Johnson study might look like Table 7.2.

A third way to present the data is in the form of a graph, which would portray the Bransford and Johnson study as you've already observed in Figure 7.12. Notice that in an experimental study, 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, however, 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 if interactions occur (Chapter 8). Tables are often preferred when there is so much data that a graph would be uninterpretable or when the researcher wishes to inform the reader of the precise values of the means; they may have to be guessed at with a graph. One rule you can certainly apply is that you should 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.

**Types of Graphs**

Notice that I have presented the Bransford and Johnson data in the form of a bar graph. Why not present it as a line graph, as in Figure 7.13? Bad idea. The problem concerns the nature of the construct being used as the independent variable and whether its underlying dimension is continuous. A **continuous variable** is one for which a number of intermediate values exist. That is, the variable exists on a continuum. An example might be the dosage level of a drug. In a study comparing 3 mg, 5 mg, and 7 mg of some drug, dosage is a continuous variable. Presumably, we also could use 4-mg or 6-mg dosages if there was a good reason for doing so. For continuous independent variables, it is appropriate to use a line graph to portray the results. That is, because it is reasonable to interpolate between the points of the graph to guess what the effects of intermediate values might be, the line can be used for estimating these in-between effects. Thus, in the drug study, a graph could look like Figure 7.14, and the researcher would be on reasonably solid ground in pre-


**Figure 7.13** Bransford and Johnson data inappropriately drawn as a line graph

dictating the effects of intermediate values of the drug, as illustrated by the point marked with an asterisk in Figure 7.15.

Of course, interpolation can be problematic if a study uses two levels of the independent variable, the two levels are far apart, and the relationship is actually nonlinear. Thus, in a drug study comparing 2-mg and 10-mg dosages that produce the solid line in Figure 7.16, interpolating the effects of a 5-mg dosage would produce a large error if the true curve looked like the dotted line. This study would be a good candidate for a single-factor multilevel design.

The situation is different if the independent variable is a discrete variable, in which each level represents a distinct category and no intermediate points can occur. In this case, no interpolation can be done, and to connect the points with a line is to imply the existence of these intermediate points when in fact they don’t exist. So when using discrete variables, as in the Bransford and Johnson study (Figure 7.12), a bar graph is normally used. The basic rule is this:

- If continuous: line graph preferred, bar graph acceptable
- If discrete: bar graphs preferred, line graph inappropriate

In general, then, bar graphs can be used for both continuous and discrete data, but line graphs should only be used for continuous data. Refer back to Box 4.3 for

**Figure 7.14** Appropriate use of a line graph with a continuous variable such as a drug's dosage level
a reminder about the ethics of presenting the data. It is easy to mislead the uninformed consumer of research by doing such things as altering the distances on the Y-axis. Your responsibility as a researcher is to present your results honestly in a way that best illustrates the true outcome of the study.

Analyzing Single-Factor, Multilevel Designs

We have seen that in the case of single-factor, two-level designs with the dependent variable measured on an interval or ratio scale, the null hypothesis can be tested with an inferential statistic called the \( t \) test. For a multilevel design such as the Bransford and Johnson 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 (e.g., context before vs. context after). Unfortunately, things 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 are the chances of having one accidentally yield significant differences. In the Bransford and Johnson study, you would have to complete 10 different \( t \) tests to cover all the pairs of conditions.

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

---

**Figure 7.15** Interpolating between the points of a line graph

**Figure 7.16** Problems with interpolation when there is a non-linear effect and a wide range between
Single Factor—More Than Two Levels

1 - (1-alpha)^c
where c = the number of comparisons being made

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

1 - (1-.05)^10 = 1 - (.95)^10 = 1 - .60 = .40

To avoid the problem of multiple t tests in single-factor designs, researchers use a procedure called a 1-way analysis of variance or 1-way ANOVA (ANalysis Of VAriance). The “1-way” means one independent variable. In essence, a 1-way ANOVA tests for the presence of some “overall” significance that could exist somewhere among the various levels of the independent variable. Hence, in a study with three levels, the null hypothesis is “level 1 = level 2 = level 3.” Repeating the null hypothesis does not identify exactly which of the “=” signs is really “≠,” however. To determine precisely where the significance lies requires what is called “subsequent testing” or “post hoc (after the fact) analysis.” 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 that some significance exists. If the ANOVA does not find any significance, subsequent testing is normally not done, unless some specific predictions about particular pairs of conditions were made ahead of time. Appendix C shows a 1-way ANOVA, followed by a common post hoc test called a “Tukey’s HSD test.”

The 1-way ANOVA yields an “F score” or an “F ratio.” Like the score from a t test, the F score examines to what extent the obtained mean differences could be due to chance or are the result of some other factor (presumably the independent variable). The ANOVA is a basic tool widely used by experimental psychologists, and if you aren’t already familiar with its operation from taking a statistics course, you should work through the examples in Appendix C. You should also be aware that even though t tests are normally used when the independent variable has just two levels, a 1-way ANOVA also could be used in that situation. Actually, the t test can be considered a special case of the ANOVA, used when there is a single independent variable with just two levels.

The designs in this chapter have in common the presence of a single independent variable. 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 with two levels of that variable. These designs can include between-subjects variables or within-subjects variables. Between-subjects variables 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 some potentially confounding variable, then randomly assigned (matched groups design). If a subject variable is used, the design is a nonequivalent groups design. Single-factor designs using a within-subjects variable are sometimes 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).
Control Group Designs
In control group designs, there is at least one condition in which the experimental treatment is absent. Varieties of control groups include placebo controls, often found in drug research; waiting list controls, found in research on the effectiveness of some type of program or therapy; and yoked controls, in which control group participants are carefully matched to treatment group participants on some factor requiring precise control.

Single Factor—More Than Two Levels
When only two levels of an experimental variable are compared, the results always will appear to be linear because a graph of the results will have only two points. Some relationships are nonlinear, however (e.g., 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 hypotheses. Like the two-level case, these multilevel designs can be either between- or within-subjects designs. Results may be presented visually through the use of bar graphs when the independent variable is a discrete variable or with a line graph if the variable is continuous. Studies using more than two levels of an independent variable are normally evaluated statistically with a 1-way analysis of variance or “ANOVA” (assuming interval or ratio data).

Chapter Review

Multiple Choice

1. Children in four different age groups (5, 7, 9, and 11 years) are tested on the Stroop task (using only the key condition that Stroop called NCWd) in order to see if experience in reading affects performance on the task (presumably, older children have more experience reading). How would you describe the design?
   a. independent groups, one-factor, multilevel
   b. single-factor, repeated-measures
   c. nonequivalent groups, one-factor, multilevel
   d. matched groups, one-factor, two-level

2. What do all single-factor, repeated-measures designs have in common?
   a. they will always have a control group
   b. participants will be tested in each condition of the study and will always be tested more than once per condition
   c. matching will be the preferred method of creating equivalent groups
   d. every participant will be tested in each of the conditions of the study

3. The case study examining the effects of sleep deprivation on being influenced by leading questions illustrated which design?
   a. within-subjects, multilevel
   b. independent groups
Chapter Review

c. matched groups
d. nonequivalent groups

4. The essential feature of a yoked control group is that
   a. its members receive a placebo
   b. what happens to its members is determined by what happens to members of the experimental group
   c. compared to experimental group members, its members receive a slightly smaller amount of the independent variable
   d. it is used whenever the researcher wishes to uncover nonlinear effects

5. Suppose you tried to replicate the Ebbinghaus memory research. It takes you 20 minutes to learn a list of CVCs on Monday and 15 minutes to relearn the list on Tuesday. What is your savings score?
   a. 25%
   b. 75%
   c. 5%
   d. cannot be determined without knowing how many items were on the list

Short Essay

1. Consider independent groups designs, matched groups designs, and nonequivalent groups designs. What do they all have in common and how do they differ?

2. In the case study that compared sleep-deprived persons with those not so deprived, why was a matched group design used, instead of an independent groups design, and what was the matching variable?

3. Describe the Stroop effect and the experimental design used by Stroop

4. Describe the two varieties of two sample t tests and with reference to the designs in the first part of the chapter (single factor—two levels), explain when each is used.

5. Use the example of the effects of alcohol on reaction time to explain the usefulness of a placebo control group

6. Use the subliminal tapes study to illustrate the usefulness of a waiting list control group.

7. Use the Weiss study of the effects of control on stress to explain what is meant by a yoked control group.

8. Use the hypothetical caffeine and reaction time study to illustrate how multilevel designs can produce nonlinear effects.

9. Use the Bransford and Johnson experiment on the effects of context on memory to illustrate the advantage of having more than two levels of an independent variable.

10. Describe when it is best to use (a) a line graph, and (b) a bar graph. Explain why a line graph would be inappropriate in a study comparing the reaction times of males and females.
Applications Exercises

Exercise 7.1.—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), and name the experimental design being used.

1. In a study of how bulimia affects the perception of body size, a group of bulimic women and a group of same-age nonbulimic women are asked to examine a graded series of 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 third is a distant city.

3. Three groups of preschoolers (50 per group, assigned randomly) are in a study of task perseverence 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 4 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 that the average verbal IQ of the groups is equivalent.

Exercise 7.2.—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 either 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 effect 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
Applications Exercises

2 months. Someone with a high level of autonomy has the ability to function well without help from others, that is, to be independent. 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 very autonomous, while R100 students are not.

3. Animals learn a maze and, in the process of so doing, 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 different 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.