9

CHAPTER

CHAPTER OUTLINE

Survey Research

Designing Your Questionnaire

Writing Questionnaire Items

Assembling Your Questionnaire

Administering Your Questionnaire

Mail Surveys

Internet Surveys

Telephone Surveys

Group-Administered Surveys

Face-to-Face Interviews

A Final Note on Survey Techniques

Assessing the Reliability of Your Questionnaire

Assessing Reliability by Repeated Administration

Assessing Reliability With a Single Administration

Increasing Reliability

Assessing the Validity of Your Questionnaire

Acquiring a Sample for Your Survey

Representativeness

Sampling Techniques

Random and Nonrandom Sampling Revisited

Sample Size

Summary

Key Terms

Using Survey Research

Gordon Allport (1954) characterized an attitude as "probably the most distinctive and indispensable concept in contemporary social psychology" (p. 43). Since Allport's assessment, attitudes have transcended social psychology to become important in our everyday lives. We are surrounded by issues related to attitudes and their measurement. Pollsters and politicians are constantly measuring and trying to change our attitudes about a wide range of issues (such as abortion, the war on terrorism, and tax cuts). How and where we obtain information on these issues are also changing.

On November 4, 2008, a historic election took place in the United States. For the first time in history an African American was elected to the office of President of the United States. Not only did the 2008 election reflect a change in America's willingness to vote for an African American candidate, it also reflected a change in how many citizens obtained their information on the candidates and the important political issues underlying the election. According to a 2009 survey conducted by the Pew Research Center, 74% of Internet users relied on the Internet to participate in or get information about the presidential election. More interestingly, there was a major increase in the percentage of adults in general as well as Internet users who obtain political news over the Internet (see Figure 9-1 for these trends).

The increased reliance on Internet sources for political news was true for a wide range of demographic groups. For example, the percentage of adults who sought political information online increased among all age groups from 2004 to 2008, with the greatest net increase among 18 to 24 year olds (a 21% increase). The increase was evident among all income groups measured (with the greatest increase among those earning less than $30,000 per year) and among Democrats (a 10% increase), Republicans (a 9% increase), and independents (a 3% increase). Additionally, the Pew survey found that Obama supporters were more likely than opponent McCain supporters to engage in a variety of online political activities. For example, Obama supporters were more likely to use social networks (25%)

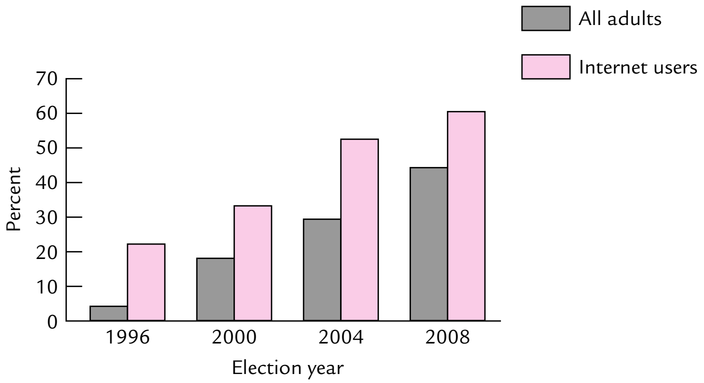


FIGURE 9-1 Trends in the use of the Internet to obtain political news.
SOURCE: http://pewresearch.org/pubs/1192/internet-politics-campaign-2008. Based on data provided at the Web site.

than McCain supporters (16%), and were more likely to post political content online (26% and 15% for Obama and McCain supporters, respectively).

Surveys are a widely used research technique. You may have participated in a survey yourself, or (perhaps more likely) you may have been the recipient of survey results. If you have answered a few questions from a local political party during election time, you have participated in a survey. Even those annoying questions on warranty registration cards that come with most products qualify as a survey of sorts. You are typically asked about your age, income, interests, magazines to which you subscribe, and so on. If you answered those questions and mailed back the card, you took part in a survey.

Even if you rarely participate in surveys, you are still likely to have encountered survey results. Political polls designed to gauge people's attitudes on key issues and candidates come out almost daily during election time. Polls about the U.S. president's approval rating, wars, and health care issues come out several times over the course of a year.

Because survey research is highly visible, you should understand the "ins and outs" of this important research technique. If you plan to use a survey technique in your own research, you should know about proper questionnaire construction, administration techniques, sampling techniques, and data analysis. Even if you never use survey techniques, understanding something about them will help you make sense out of the surveys that you are exposed to every day.

SURVEY RESEARCH

Before we discuss survey techniques, note the difference between the *field survey* and the observational techniques described in Chapter 8. In both naturalistic observation and participant observation, you simply observe behaviors and make copious notes about them. You do not administer any measures to your participants. Consequently, you can only speculate about the motives, attitudes, and beliefs underlying the observed behaviors. In a field survey, you directly question your participants about their behavior (past, present, or future) and their underlying attitudes, beliefs, and intentions. From the data collected, you can draw inferences about the factors underlying behavior.

The inferences that you can draw from a field survey are limited by the fact that you do not manipulate independent variables. Instead, you acquire several (perhaps hundreds of) measures of the behaviors of interest. This purely correlational research strategy usually does not permit you to draw causal inferences from your data (see Chapter 4). For example, finding that political conservatism is a good predictor of voter choices does not justify concluding that political conservatism *causes* voter choices.

Instead, you use the field survey to evaluate specific attitudes such as those concerning issues surrounding nuclear disarmament, political candidates, or foreign imports. You also can use the field survey to evaluate behaviors. For example, you could design a questionnaire to determine which household products people use.

Surveys also have another important use: predicting behavior. Political polls often seek to predict behavior. Attitudes about political candidates are assessed, and then projections are made about subsequent voter behavior.

When you conduct survey research, you must ensure that your participants are treated ethically. One major ethical issue concerns whether and how you will maintain the *anonymity* of your participants and the *confidentiality* of their responses. Maintaining anonymity means that you guarantee there will be no way for the participants' names to be associated with their answers. This might be accomplished by instructing participants to mail back their questionnaires and informed-consent forms separately. No coding scheme would be used that would allow you to match up individual participants and their questionnaires. However, sometimes you may wish to code the questionnaires and informed-consent forms so that you can match them up later. You might do this, for example, if a participant has second thoughts about participating after the questionnaire has been returned. If so and you have promised your participants that their responses will remain anonymous, you must take steps to ensure that only authorized personnel associated with the research project can gain access to the code and only for the stated purpose.

Maintaining confidentiality means that you do not disclose any data in individual form, even if you know which participants filled out which questionnaires. If you promise your participants that their responses will remain confidential, ethical practice dictates that you report only aggregate results.

QUESTIONS TO PONDER

1. What are some of the applications of survey research?
2. Why is it important to know about survey methods, even if you do not intend to conduct surveys?
3. How does a field survey differ from other observational methods?
4. What are anonymity and confidentiality and why are they important?

DESIGNING YOUR QUESTIONNAIRE

The first step in designing a questionnaire is to clearly define the topic of your study. A clear, concise definition of what you are studying will yield results that can be interpreted unambiguously. Results from surveys that do not clearly define the topic area may be confusing. It is also important to have clear, precise operational definitions for the attitudes or behaviors being studied. Behaviors and attitudes that are not defined precisely also may yield results that are confusing and difficult to interpret.

Having a clearly defined topic has another important advantage: It keeps your questionnaire focused on the behavior or attitude chosen for study (Moser & Kalton, 1972). You should avoid the temptation to do too much in a single survey. Tackling too much in a single survey leads to an inordinately long questionnaire that may confuse or overburden your participants. It also may make it more difficult for you to summarize and analyze your data (Moser & Kalton, 1972). Your questionnaire should include a broad enough range of questions so that you can thoroughly assess behavior but not so broad as to lose focus and become confusing. Your questionnaire should elicit the responses you are most interested in without much extraneous information.

The type of information gathered in a questionnaire depends on its purpose. However, most questionnaires include items designed to assess the characteristics of the participants, such as age, sex, marital status, occupation, income, and education. Such characteristics are called *demographics*. Demographics are often used as *predictor variables* during analysis of the data to determine whether participant characteristics correlate with or predict responses to other items in the survey. Other, nondemographic items also can be included to provide predictor variables. For example, attitude toward abortion might be used to predict voter preference. In this case, attitude toward abortion would be used as a predictor variable.

In addition to demographics and predictor variables, you will have items designed to assess the behavior of interest. For example, if you were interested in predicting voter preference, you would include an item or items on your questionnaire specifically to measure voter preference (e.g., asking participants to indicate candidate preferences). That item, or a combination of several items, would constitute the *criterion variable*.

The questions to which your participants will respond are the heart of your questionnaire. Take great care to develop questions that are clear, to the point, and relevant to the aims of your research. The time spent in this early phase of your research will pay dividends later. Well-constructed items are easier to summarize, analyze, and interpret than poorly constructed ones. The next section introduces several popular item formats and offers suggestions for writing good questionnaire items.

Writing Questionnaire Items

Writing effective questionnaire items that obtain the information you want requires care and skill. You cannot simply sit down, write several questions, and use those first-draft questions on your final questionnaire. Writing questionnaire items involves writing and rewriting items until they are clear and succinct. In fact, having written your items and assembled your questionnaire, you should administer it to a pilot group of participants matching your main sample in order to ensure that the items are reliable and valid.

When writing questionnaire items, you may choose among several popular types. Here we discuss the open-ended, restricted, partially open-ended, and rating-scale item types.

Open-Ended Items **Open-ended items** allow the participant to respond in his or her own words. The following example might appear in a survey like the Pew Internet use survey:

> **How often did you use the Internet to get political news for the 2008 presidential election?**

The participant writes an answer to the question in the space provided immediately below. Such information may be more complete and accurate than the information obtained with a restricted item (discussed next). A drawback to the open-ended item is that participants may not understand exactly what you are looking for or may inadvertently omit some answers. Thus, participants may fail to provide the needed information. Another drawback to the open-ended item is that it can make summarizing your data difficult. Essentially, you must perform a content analysis on open-ended answers. All of the methods and rules that we discussed in Chapter 8 would come into play. It may be tempting to interpret open-ended responses rather than just summarize them, running the risk of misclassifying the answers.

Restricted Items **Restricted items** (also called *closed-ended items*) provide a limited number of specific response alternatives. A restricted item with ordered alternatives lists these alternatives in a logical order, as shown in this item adapted from the Pew survey:

> **How often did you use the Internet to get political news during the 2008 presidential election campaign?**
>
> __ Very often
>
> __ Sometimes
>
> __ Not too often
>
> __ Never

Note how the alternatives for this question go from very often to never. Participants would respond by checking the blank space to the left of the desired answer. However, other methods for recording choices can be used with restricted items. For example, you could use a number to the right of each alternative and have participants circle the numbers corresponding to their choices.

Use unordered alternatives whenever there is no logical basis for choosing a given order, as shown in this example from the Pew survey:

Do you think that the political information you obtained from the Internet during the 2008 presidential election campaign was generally accurate or inaccurate?

__ Accurate

__ Inaccurate

__ Neither

__ Don't know

Because there is no inherent order to the alternatives, other orders would serve just as well. For example, you just as easily could have put "Inaccurate" before "Accurate."

By offering only specific response alternatives, restricted items control the participant's range of responses. The responses made to restricted items are therefore easier to summarize and analyze than the responses made to open-ended items. However, the information that you obtain from a restricted item is not as rich as the information from an open-ended item. Participants cannot qualify or otherwise elaborate on their responses. Also, you may fail to include an alternative that correctly describes the participant's opinion, thus forcing the participant to choose an alternative that does not really fit.

Partially Open-Ended Items **Partially open-ended items** resemble restricted items but provide an additional, "other" category and an opportunity to give an answer not listed among the specific alternatives, as shown in this example adapted from the Pew survey:

In what capacity did you most use the Internet during the 2008 presidential election campaign?

__ Post political content online

__ Engage politically on an online social network

__ Share political videos, pictures, or audio content

__ Sign up for online political updates

__ Donate money online

__ Other (Specify) ____________________

Dillman (2000) offers several suggestions for formatting restricted and partially open-ended items. First, use a boldface font for the stem of a question and a normal font for response category labels (as we have done in the previous examples). This helps respondents separate the question from the response categories that follow. Second, make any special instructions intended to clarify a question a part of the question itself. Third, put check boxes, blank spaces, or numbers in a consistent position throughout your questionnaire (e.g., to the left of the response alternatives). Fourth, place all alternatives in a single column. Other tips offered by Dillman (2000) for constructing and formatting questionnaire items are summarized in Table 9-1.

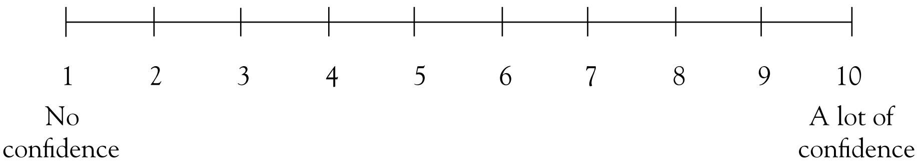
TABLE 9-1 Suggestions for Writing Good Survey Items

SUGGESTION	EXAMPLE
Use simple rather than complex words.	Use "work" rather than "employment."
Make the stem of a question as short and easy to understand as possible, but use complete sentences.	"Would you like to study in America?"
Avoid vague questions in favor of more precise ones.	Use "How many years have you lived in your current house?" rather than "Years in your house."
Avoid asking for too much information. Respondents may not have an answer readily available.	Use a list of ordered alternatives rather than an open-ended question when asking how often the respondent does something.
Avoid "check all that apply" questions.	Instead of "check all that apply," list each item separately and have respondent indicate liking/disliking for each.
Avoid questions that ask for more than one thing.	Instead of asking "Would you like to study and then live in America?" ask "Would you like to study in America?" and "Would you like to live in America?" separately.
Soften the impact of potentially sensitive questions.	Instead of asking "Have you ever stolen anything?" ask "Have you ever taken anything without paying for it?"

SOURCE: After Dillman, 2000.

Rating Scales A variation on the restricted question uses a rating scale rather than response alternatives. A rating scale provides a graded response to a question:

> **How much confidence do you have that the political news you obtained from the Internet during the 2008 presidential campaign was accurate?**



There is no set number of points that a rating scale must have. A rating scale can have as few as 3 and as many as 100 points. However, rating scales commonly do not exceed 10 points. A 10-point scale has enough points to allow a wide range of choice while not overburdening the participant. Scales with fewer than 10 points also are used frequently, but you should not go below 5 points. Many participants may not want to use the extreme values on a scale. Consequently, if you have a 5-point scale and the participant excludes the end points, you really have only three usable points. Scales ranging from 7 to 10 points leave several points for the participants to choose among, even if participants do avoid the extreme values.

You also must decide how to label your scale. Figure 9-2 shows three ways that you might do this. In panel (a), only the end points are labeled. In this case, the participant is told the upper and lower limits of the scale. Such labeled points are called *anchors* because they keep the participant's interpretation of the scale values from drifting.

With only the end points anchored, the participant must interpret the meaning of the rest of the points. In Figure 9-2(b), all points are labeled. In this case, the participant knows exactly what each point means and may consequently provide more accurate information. In Figure 9-2(c), the scale is labeled at the end points and at the midpoint. This scale provides three anchors for the participant. This scale is a reasonable compromise between labeling only the end points and labeling all the points.

You may be wondering whether labeling each point changes the way that the participant responds on the scale. The answer seems to be a qualified no. When you develop a measurement scale, you are dealing with (1) the psychological phenomenon underlying the scale and (2) the scale itself. Labeling each point does not change the nature of the psychological phenomenon underlying the scale. You can assume that your scale, labeled at each point, still represents the phenomenon underlying the scale. In fact, researchers have sometimes expressed a misguided concern about such scale transformations (Nunnally, 1967). Minor transformations of a measurement

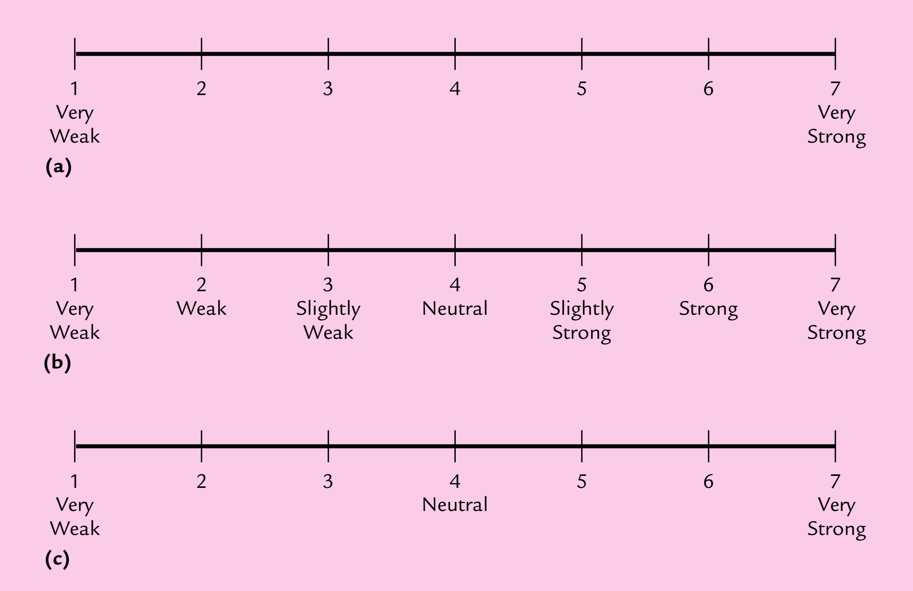


FIGURE 9-2 Three ways of labeling a rating scale: (a) end points only, (b) each point labeled, and (c) end points and midpoint labeled.

scale (such as labeling each point) probably do not affect its measurement properties or how well it represents the underlying psychological phenomenon being studied.

In the previous examples, participants respond by checking or circling the scale value that best represents their judgments. Alternative ways to format your scale give participants more flexibility in their responses. Figure 9-3 shows an example in which the end points are anchored and the participants are instructed to place a check or perpendicular line on the scale to indicate how they feel. To quantify the responses, you use a ruler to measure from an end point to the participant's mark. Your scale is then expressed in terms of inches or centimeters, and the resulting numbers are treated just like the numbers on a numbered scale.

Another variation on the rating scale is the *Likert scale*, which is widely used in attitude measurement research. A Likert scale provides a series of statements to which participants can indicate degrees of agreement or disagreement. Figure 9-4 shows two examples of formatting a Likert-scale item. In the first example, the attitude statement is followed by five blank spaces labeled from "Strongly Agree" to "Strongly Disagree." The participant simply checks the space that best reflects the degree of agreement or disagreement with each statement. The second example provides consecutive numbers rather than blank spaces and includes descriptive anchors only at the ends. Participants are instructed to circle the number that best reflects how much they agree or disagree with each statement. (For further information on Likert scaling, see Edwards, 1953).

A final note on rating scales is in order. Although rating scales have been presented in the context of survey research, be aware that rating scales are widely used in experimental research as well. Adapting rating scales to your particular research needs is a relatively simple affair. Anytime that your research calls for the use of rating scales, you can apply the suggestions presented here.

QUESTIONS TO PONDER

1. What are the steps involved in designing a questionnaire?
2. How do open-ended and restricted items differ, and what are the advantages and disadvantages of each?
3. What are the ways in which questionnaire items can be formatted?
4. What are some of the factors that you should pay attention to when constructing questionnaire items?
5. How do you design effective rating scales?

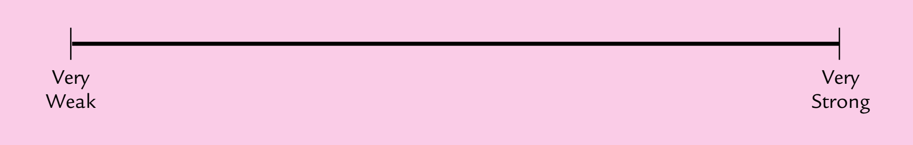


FIGURE 9-3 Rating scale formatted with no numbers. End points are labeled, and participants place marks on the line to indicate their responses.

Most political information on the Internet is accurate.

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
________	________	________	________	________

(a)

Most political information on the Internet is accurate.

Strongly Agree				Strongly Disagree
1	2	3	4	5

(b)

FIGURE 9-4 Samples showing Likert scales: (a) a standard Likert item on which the participant places a check in the blank under the statement that best reflects how he or she feels; (b) a five-point Likert scale using numbers that the participant circles.

Assembling Your Questionnaire

If your questionnaire is to be effective, its items must be organized into a coherent, visually pleasing format. This process involves paying attention to the order in which the items are included and to the way in which they are presented.

Dillman (2000) and Moser and Kalton (1972) agree that demographic items should *not* be presented first on the questionnaire. These questions, although easy to complete, may lead participants to believe that the questionnaire is boring. Dillman emphasizes the importance of the first question on a questionnaire. A good first question should be interesting and engaging so that the respondent will be motivated to continue. According to Dillman, the first question should apply to everybody completing the questionnaire, be easy so that it takes only a few seconds to answer, and be interesting. Of course, these rules are not carved in stone. If your research needs require a certain question to be presented first, that consideration should take precedence (Dillman, 2000).

Your questionnaire should have continuity; that is, related items should be presented together. This keeps your participant's attention on one issue at a time rather than jumping from issue to issue. Your questionnaire will have greater continuity if related items are grouped. An organized questionnaire is much easier and more enjoyable for the participant to complete, factors that may increase the completion rate. Continuity also means that groups of related questions should be logically ordered. Your questionnaire should read like a book. Avoid the temptation to skip around from topic to topic in an attempt to hold the attention of the participant. Rather, strive to build "cognitive ties" between related groups of items (Dillman, 2000).

The order in which questions are included on a questionnaire has been shown to affect the responses of participants. For example, McFarland (1981) presented questions on a questionnaire ordered in two ways. Some participants answered a general question before specific questions, whereas others answered the specific questions first. McFarland found that participants expressed more interest in politics and religion when the specific questions were asked first than when the general questions were asked first. Sigelman (1981) found that question order affected whether or not participants expressed an opinion (about the popularity of the president), but only if the participants were poorly educated. Hence, question order may play a greater role for some participants than for others. Carefully consider your sample and the chosen topic when deciding on the order in which questions are asked.

The placement of items asking for sensitive information (such as sexual preferences or illegal behavior) is an important factor. Dillman (2000) suggests placing objectionable questions after less objectionable ones, perhaps even at the end of the questionnaire. Once your participants are committed to answering your questions, they may be more willing to answer some sensitive questions. Additionally, a question may not seem as objectionable after the respondent has answered previous items than if the objectionable item is placed earlier in the questionnaire (Dillman, 2000). You also should pay attention to the way that each page of your questionnaire is set up. There should be a logical "navigational path" (Dillman, 2000) that your respondent can follow. This path should lead the respondent through the questionnaire as if he or she were reading a book.

One way to accomplish this is to use appropriate graphics (e.g., arrows and other symbols) to guide respondents through the questionnaire. In fact, Dillman talks about two "languages" of a questionnaire. One language is verbal and relates to how your questions are worded. The other language is graphical and relates to the symbols and graphics used to guide respondents through the items on your questionnaire. Symbols and graphics can be used to separate groups of items, direct respondents where to go in the event of a certain answer (e.g., "If you answered 'No' to item 5, skip to item 7" could be accompanied by an arrow pointing to item 7), or direct respondents to certain pages on the questionnaire. Dillman suggests the following three steps for integrating the verbal and graphical languages into an effective questionnaire:

1. Design a navigational path directing respondents to read all the information on a page.
2. Create effective visual navigational guides to help respondents stay on the navigational path.
3. Develop alternate navigational guides to help with situations where the normal navigational guide will be interrupted (e.g., skipping items or sections).

QUESTIONS TO PONDER

1. Why is the first question on a questionnaire so important?
2. What does it mean that a questionnaire should have continuity? Why is continuity important?
3. What is a questionnaire's navigational path, and why is it important?

ADMINISTERING YOUR QUESTIONNAIRE

After you develop your questionnaire, you must decide how to administer it. You could mail your questionnaire to your participants, deliver your questionnaire via e-mail or post it on the Internet, telephone participants to ask the questions directly, administer your questionnaire to a large group at once, or conduct face-to-face interviews. Each method has advantages and disadvantages and makes its own special demands.

Mail Surveys

In a **mail survey**, you mail your questionnaire directly to your participants. They complete and return the questionnaire at their leisure. This is a rather convenient method. All you need to do is put your questionnaires into addressed envelopes and mail them. However, a serious problem called **nonresponse bias** occurs when a large proportion of participants fail to complete and return your questionnaire. If the participants who fail to return the questionnaire differ in significant ways from those who do return it, your survey may yield answers that do not represent the opinions of the intended population.

Combating Nonresponse Bias To reduce nonresponse bias, you should develop strategies to increase your return rate. Dillman (2000) notes that the single most effective strategy for increasing response rate is to make multiple contacts with respondents. Dillman suggests making four contacts via mail. The first consists of a *prenotice letter* sent to the respondent a few days before the questionnaire is sent. The prenotice letter should inform the respondent that an important questionnaire will be coming in the mail in a few days. It also should inform the respondent what the survey is about and why the survey will be useful. The second mailing would deliver the questionnaire itself, accompanied by a cover letter. The cover letter should include the following elements in the order listed (Dillman, 2000): the specific request to complete the questionnaire, why the respondent was selected to receive the survey, the usefulness of the survey, a statement of confidentiality of the respondent's answers, an offer of a token of appreciation (if such an offer is to be made), an offer to answer questions, and a real signature.

The third mailing would take the form of a *thank you postcard* sent a few days or a week after the questionnaire was mailed. The postcard should thank the respondent for completing the questionnaire and remind the respondent to complete the questionnaire if not already done. The fourth contact provides a *replacement questionnaire*, sent 2 to 4 weeks after the original questionnaire and accompanied by a letter indicating that the original questionnaire had not been received. The letter also should urge the respondent to complete the replacement questionnaire and return it.

You may be able to increase your return rate somewhat by including a small token of your appreciation, such as a pen or pencil that the participant can keep. Some researchers include a small amount of money as an incentive to complete the questionnaire. As a rule, it is better to send the token along with the questionnaire rather than make the token contingent upon returning the questionnaire. One study found that 57% of respondents returned a survey questionnaire when promised $50 for its return whereas 64% returned the questionnaire when $1 was included with it (James & Bolstein, 1990).

Ironically, smaller rewards seem to produce better results than larger ones (Kanuk & Berenson, 1975; Warner, Berman, Weyant, & Ciarlo, 1983). Dillman (2000) suggests that a $1 token is preferred because it is easy to mail and seems to produce the desired results. Finally, monetary incentives work better than tangible rewards (Church, 1993).

A few factors that do *not* significantly affect response rate include questionnaire length, personalization, promise of anonymity, and inclusion of a deadline (Kanuk & Berenson, 1975). (For reviews of the research supporting these findings, see Kanuk & Berenson, 1975, and Warner et al., 1983.)

Internet Surveys

An increasingly popular method of administering questionnaires is to post them on the Internet. **Internet surveys** can be distributed via e-mail or listserves or posted on a Web site. Which method you use depends on the nature and purpose of your survey. E-mail surveys are easy to distribute but do not permit complex navigational designs (Dillman, 2000). Consequently, e-mail surveys are best for relatively short, simple questionnaires. Web-based surveys allow you to create and present more complex questionnaires that incorporate many of the design features discussed previously (Dillman, 2000). To aid you in the task of implementing a Web-based survey, commercial software packages are available that allow you to design sophisticated questionnaires for posting on a Web site. There is significant advantage to using the Internet to conduct a survey or recruit participants: You can reach a large body of potential participants with relative ease. Data can be collected quickly and easily, resulting in a large data set. You still need to consider the problem of nonresponse bias. As with the mail survey, you can combat this problem with prenotification. For an Internet survey a short text message to potential respondents is more effective than an e-mail notice (Bosnjak, Neubarth, Couper, Bandilla, & Kaczmirek, 2008).

There are also disadvantages to Internet surveys. As discussed in Chapter 6, a sample of respondents from the Internet may not be representative of the general population. According to a 2007 study by the U.S. Department of Commerce (2008), only 61.7% of households had access to the Internet in the home. Further, households with higher levels of education and income were more likely to have Internet access. Additionally, access was greater for Asians (75.5%) and Whites (67.0%) than Blacks (44.9%). Another disadvantage is that one must have the resources available to post a survey on the Internet. This requires computer space on a server and the ability to create the necessary Web pages or the resources to pay someone to create your net survey for you.

Despite the potential for biased samples in Internet surveys, there is evidence that the results obtained from Internet surveys are equivalent to the results obtained from paper-and-pencil surveys. Alan De Beuckelear and Flip Lievens (2009) conducted a survey across 16 countries using both Internet and paper-and-pencil deliveries. The results showed that in all of the countries the Internet and paper-and-pencil surveys returned equivalent results. De Beuckelear and Lievens (2009) concluded that data collected with the two methods could be combined because the two methods produced such highly similar data. In another study, Christopher Fleming and Mark Bowden (2009) found that the sample demographics of an Internet and a mail survey on travel preferences did not differ significantly.

In both of the studies just cited, the topics of the surveys were not sensitive or controversial. There is some evidence that the equivalence of Internet and conventional methods may not apply to more sensitive topics (DiNitto, Busch-Armendariz, Bender, Woo, Tackett-Gibson, & Dyer, 2009). DiNitto, et al. conducted a survey over the Internet and by telephone asking men about sexual assault behaviors. The results showed that respondents in both types of survey reported sexual assault behavior. However, a wider variety of sexual assault behaviors were reported by respondents to the telephone survey.

So, where does this leave us? It would appear that Internet surveys may produce comparable results to other survey methods for nonsensitive issues. You can be reasonably confident that your Internet survey on such issues will yield data that are highly similar to data collected with more conventional methods. However, you must exercise more caution when surveying about sensitive behaviors. In the latter case, an Internet survey may produce results that differ from more conventional methods.

Telephone Surveys

In a **telephone survey**, you contact participants by telephone rather than by mail or via the Internet. You can ask some questions more easily over the telephone than you can in written form. Telephone surveys can be done by having an interviewer ask respondents a series of questions or by interactive voice response (IVR). Telephone surveys using live interviewers have lost popularity as new technologies have become available. IVR surveys involve respondents using a touch-tone telephone to respond to a series of prerecorded questions. Modern IVR technologies also allow respondents to provide verbal answers in addition to numeric responses.

Telephone surveys may not be the best way to administer a questionnaire. The plethora of "junk calls" to which the population is exposed has given rise to a backlash against telephone intrusions. Laws have been passed on the state and federal level protecting people from unwanted calls, making it more difficult to reach prospective respondents. These laws, combined with caller ID and answering machines (which allow residents to screen their calls), make the telephone a less attractive medium for surveys now than in the past.

Group-Administered Surveys

Sometimes you may have at your disposal a large group of individuals to whom you can administer your questionnaire. In such a case, you design your questionnaire as you would for a mail survey but administer it to the assembled group. For example, you might distribute to a first-year college class a questionnaire on attitudes toward premarital sex. Using such a captive audience permits you to collect large amounts of data in a relatively short time. You do not have to worry about participants misplacing or forgetting about your questionnaire. You also may be able to reduce any volunteer bias, especially if you administer your questionnaire during a class period. People may participate because very little effort is required.

As usual, this method has some drawbacks. Participants may not treat the questionnaire as seriously when they fill it out as a group as when they fill it out alone. Also, you may not be able to ensure anonymity in the large group if you are asking for sensitive information. Participants may feel that other participants are looking at their answers. (You may be able to overcome this problem by giving adjacently seated participants alternate forms of the questionnaire.) Also, a few participants may express hostility about the questionnaire by purposely providing false information.

A final drawback to group administration concerns the participant's right to decline participation. A participant may feel pressure to participate in your survey. This pressure arises from the participant's observation that just about everyone else is participating. In essence, a conformity effect occurs because completing your survey becomes the norm defined by the behavior of your other participants. Make special efforts to reinforce the understanding that participants should not feel compelled to participate.

Face-to-Face Interviews

Still another method for obtaining survey data is the **face-to-face interview**. In this method, you talk to each participant directly. This can be done in the participant's home or place of employment, in your office, or in any other suitable place. If you decide to use a face-to-face interview, keep several things in mind. First, decide whether to use a structured interview or an unstructured interview. In a *structured interview,* you ask prepared questions. This is similar to the telephone survey in that you prepare a questionnaire in advance and simply read the ordered questions to your participants. In the *unstructured interview,* you have a general idea about the issues to discuss. However, you do not have a predetermined sequence of questions.

An advantage of the structured interview is that all participants are asked the same questions in the same order. This eliminates fluctuations in the data that result from differences in when and how questions are asked. Responses from a structured interview are therefore easier to summarize and analyze. However, the structured interview tends to be inflexible. You may miss some important information by having a highly structured interview. The unstructured interview is superior in this respect. By asking general questions and having participants provide answers in their own words, you may gain more complete (although perhaps less accurate) information. However, responses from an unstructured interview may be more difficult to code and analyze later on. You can gain some advantages of each method by combining them in one interview. For example, begin the interview with a structured format by asking prepared questions; later in the interview, switch to an unstructured format.

Using the face-to-face interview strategy leads to a problem that is not present in mail or Internet surveys but is present to some extent in telephone surveys: The appearance and demeanor of the interviewer may affect the responses of the participants. Experimenter bias and demand characteristics become a problem. Subtle changes in the way in which an interviewer asks a question may elicit different answers. Also, your interviewer may not respond similarly to all participants (e.g., an interviewer may react differently to an attractive participant than to an unattractive one). This, too, can affect the results.

The best way to combat this problem is to use interviewers who have received extensive training in interview techniques. Interviewers must be trained to ask questions in the same way for each participant. They also must be trained not to emphasize any particular words in the stem of a question or in the response list. The questions should be read in a neutral manner. Also, try to anticipate any questions that participants may have and provide your interviewers with standardized responses. This can be accomplished by running a small pilot version of your survey before running the actual survey. During this pilot study, try out the interview procedure on a small sample of participants. (This can be done with just about anyone, such as friends, colleagues, or students.) Correct any problems that arise.

Another problem with the interview method is that the social context in which the interview takes place may affect a participant's responses. For example, in a survey of sexual attitudes known as the "Sex in America" survey (Michael, Gagnon, Laumann, & Kolata, 1994), some questions were asked during a face-to-face interview. Some participants were interviewed alone whereas others were interviewed with a spouse or other sex partner present. Having the sex partner present changed the responses to some questions. For example, when asked a question about the number of sex partners one had over the past year, 17% of the participants interviewed alone reported two or more. When interviewed with their sex partner present, only 5% said they had two or more sex partners. It would be most desirable to conduct the interviews in a standardized fashion with only the participant present.

A Final Note on Survey Techniques

Although each of the discussed techniques has advantages, the mail survey has been the most popular. The mail survey can reach large numbers of participants at a lower cost than either the telephone survey or the face-to-face interview (Warner et al., 1983) and produces data that are less affected by *social desirability effects* (answering in a way that seems socially desirable). For these reasons, consider mail surveys first.

After designing your questionnaire and choosing a method of administration, the next step is to assess the reliability and validity of your questionnaire. This is typically done by administering your questionnaire to a small but representative sample of participants. Based on the results, you may have to rework your questionnaire to meet acceptable levels of reliability and validity. In the next sections, we introduce you to the processes of evaluating the reliability and validity of your questionnaire.

QUESTIONS TO PONDER

1. What are the different ways of administering a questionnaire?
2. What are the advantages and disadvantages of the different ways of administering a questionnaire?
3. What is nonresponse bias and what can you do to combat it?
4. How do social desirability effects affect your decision about how to administer a questionnaire?

ASSESSING THE RELIABILITY OF YOUR QUESTIONNAIRE

Constructing a questionnaire is typically not a one-shot deal. That is, you don't just sit down and write some questions and magically produce a high-quality questionnaire. Developing a quality questionnaire usually involves designing the questionnaire, administering it, and then evaluating it to see if it does the job.

One dimension you must pay attention to is the reliability of your questionnaire. In Chapter 5, we defined *reliability* as the ability of a measure to produce the same or highly similar results on repeated administrations. This definition extends to a questionnaire. If, on testing and retesting, your questionnaire produces highly similar results, you have a reliable instrument. In contrast, if the responses vary widely, your instrument is not reliable (Rogers, 1995).

In Chapter 5, we described two ways to assess the reliability of a measure: the test–retest method and the split-half method. In the next sections, we discuss the application of these two methods when assessing the reliability of a questionnaire.

Assessing Reliability by Repeated Administration

Evaluating test–retest reliability is the oldest and conceptually simplest way of establishing the reliability of your questionnaire. You simply administer your questionnaire, allow some time to elapse, and then administer the questionnaire (or a parallel form of it) again to the same group of participants. Although this method is relatively simple to execute, you need to consider some issues before using it.

First, you must consider how long to wait between administrations of your questionnaire. An intertest interval that is too short may result in participants remembering your questions and the answers they gave. This could lead to an artificially high level of test–retest reliability. If, however, you wait too long, test–retest reliability may be artificially low. According to Tim Rogers (1995), the intertest interval should depend on the nature of the variables being measured, with an interval of a few weeks being sufficient for most applications. Rogers suggests that test–retest methods may be particularly problematic when applied to the following:

1. *Measuring ideas that fluctuate with time.* For example, an instrument to measure attitudes toward universal health care should not be evaluated with the test–retest method because attitudes on this topic seem to shift quickly.
2. *Issues for which individuals are likely to remember their answers on the first testing.*
3. *Questionnaires that are very long and boring.* The problem here is that participants may not be highly motivated to accurately complete an overly long questionnaire and therefore may give answers that reduce reliability.

Some of the problems inherent in using the *same* measure on multiple occasions can be avoided by using alternate or parallel forms of your questionnaire for multiple testing sessions. As noted in Chapter 5, the type of reliability being assessed with this technique is known as parallel-forms reliability (Rogers, 1995).

For the parallel-forms method to work, the two (or more) forms of your questionnaire must be equivalent so that direct comparison is meaningful. According to Rogers (1995), parallel forms should have the same number of items and the same response format, cover the same issues with different items, be equally difficult, use the same instructions, and have the same time limits. In short, the parallel versions of a test must be as equivalent as possible (Rogers, 1995).

Although the parallel-forms method improves on the test–retest method, it does not solve all the problems associated with multiple testing. Using parallel forms does not eliminate the possibility that rapidly changing attitudes will result in low reliability. As with the test–retest method, such changes make the questionnaire appear less reliable than it actually is. In addition, practice effects may occur even when alternate forms are used (Rogers, 1995). Even though you use different questions on the parallel form, participants may respond similarly on the second test because they are familiar with your question format.

Assessing Reliability With a Single Administration

Because of the problems associated with repeated testing, you might consider assessing reliability by means of a single administration of your questionnaire. As noted in Chapter 5, this approach involves splitting the questionnaire into equivalent halves and deriving a score for each half; the correlation between scores from the two halves is known as split-half reliability (Rogers, 1995). This technique works best when your survey is limited to a single specific area (e.g., sexual behavior) as opposed to multiple areas (sexual behavior and sexual attitudes).

Although the split-half method circumvents the problems associated with repeated testing, it introduces others. First, when you split a questionnaire, each score is based on a limited set of items, which can reduce reliability (Rogers, 1995). Consequently, the split-half method may underestimate reliability. Second, it is not clear how splitting should be done. If you simply do a first-half/second-half split, artificially low reliability may occur if the two halves of the form are not equivalent or if participants are less motivated to answer questions accurately on the second half of your questionnaire and therefore give inconsistent answers to your questions. One remedy for this is to use an odd–even split. In this case, you derive a score for the odd items and a score for the even items.

Perhaps the most desirable way to assess the split-half reliability of your questionnaire is to apply the Kuder–Richardson formula. This formula yields the average of all the split-half reliabilities that could be derived from splitting your questionnaire into two halves in every possible way. The resulting number (designated KR20) will lie between 0 and 1; the higher the number, the greater the reliability of your questionnaire. A KR20 of .75 indicates a "moderate" level of reliability (Rogers, 1995).

In cases in which your questionnaire uses a Likert format, a variation on the Kuder–Richardson formula known as *coefficient alpha* is used (Rogers, 1995). Like KR20, coefficient alpha is a score between 0 and 1, with higher numbers indicating greater reliability. Computation of this formula can be complex. For details, see a text on psychological testing (e.g., see Cohen & Swerdlik, 2010; Rogers, 1995).

Increasing Reliability

Regardless of the method you use to assess the reliability, there are steps you can take to increase the reliability of your questionnaire (Rogers, 1995):

1. Increase the number of items on your questionnaire. Generally, higher reliability is associated with increasing numbers of items. Of course, if your instrument becomes too long, participants may become angry, tired, or bored. You must weigh the benefits of increasing questionnaire length against possible liabilities.
2. Standardize administration procedures. Reliability will be enhanced if you treat all participants alike when administering your questionnaire. Make sure that timing procedures, lighting, ventilation, instructions to participants, and instructions to administrators are kept constant.
3. Score your questionnaire carefully. Scoring errors can reduce reliability.
4. Make sure that the items on your questionnaire are clear, well written, and appropriate for your sample (see our previous discussion on writing items).

QUESTIONS TO PONDER

1. What is meant by the reliability of a questionnaire and why is it important?
2. How do you assess reliability with repeated administrations?
3. How do you assess reliability with a single administration?
4. What steps can be taken to increase reliability?

ASSESSING THE VALIDITY OF YOUR QUESTIONNAIRE

In Chapter 5, we discussed the validity of a measure and described several forms of validity that differ in their method of assessment: content validity, criterion-related validity, construct validity, and face validity. As with other measures, a questionnaire must have validity if it is to be useful; that is, it must measure what it is intended to measure. For example, if you are designing a questionnaire to assess political attitudes, the questions on your test should tap into political attitudes and not, say, religious attitudes.

Here we review content validity, construct validity, and criterion-related validity as applied to a questionnaire (Rogers, 1995). In a questionnaire, *content validity* assesses whether the questions cover the range of behaviors normally considered to be part of the dimension that you are assessing. To have content validity, your questionnaire on political attitudes should include items relevant to all the major issues relating to such attitudes (e.g., abortion, health care, the economy, and defense). The *construct validity* of a questionnaire can be established by showing that the questionnaire's results agree with predictions based on theory.

Establishing the *criterion-related validity* of a questionnaire involves correlating the questionnaire's results with those from another, established measure. There are two ways to do this. First, you can establish *concurrent validity* by correlating your questionnaire's results with those of another measure of the *same* dimension administered at the same time. In the case of your questionnaire on political attitudes, you would correlate its results with those of another, established measure of political attitudes. Second, you can establish *predictive validity* by correlating the questionnaire's results with some behavior that would be expected to occur, given the results. For example, your questionnaire on political attitudes would be shown to have predictive validity if the questionnaire's results correctly predicted election outcomes.

The validity of a questionnaire may be affected by a variety of factors. For example, as noted earlier, how you define the behavior or attitude that you are measuring can affect validity. Validity also can be affected by the methods used to gather your data. In the "Sex in America" survey, some respondents were interviewed alone and others with someone else present. One cannot be sure that the responses given with another person present represent an accurate reflection of one's sexual behavior (Stevenson, 1995). Generally, methodological flaws, poor conceptualization, and unclear questions can all contribute to lowered levels of validity.

QUESTIONS TO PONDER

1. What is the validity of a questionnaire and why is it important?
2. What are the different types of validity you should consider?
3. What factors can affect the validity of your questionnaire?

ACQUIRING A SAMPLE FOR YOUR SURVEY

In Chapter 6, we distinguished between a population (all individuals in a well-defined group) and a sample (a smaller number of individuals selected from the population). Once you have designed and pretested your questionnaire, you then administer it to a group of participants. It is usually impractical to have everyone in the population (however that may be defined) complete your survey. Instead, you administer your questionnaire to a small sample of that population.

Proper sampling is a crucial aspect of sound survey research methodology. Without proper sampling, you can't generalize your results to your target population (e.g., accurately predict voter behavior in an election). Three sampling-related issues you must consider are representativeness, sampling technique, and sample size.

Representativeness

Regardless of the technique you use to acquire your sample, your sample should be representative of the population of interest. A **representative sample** closely matches the characteristics of the population. Imagine that you have a bag containing 300 golf balls: 100 are white, 100 are orange, and 100 are yellow. You then select a sample of

30 golf balls. A representative sample would have 10 balls of each color. A sample having 25 white and 5 orange would not be representative (the ratio of colors does not approximate that of the population) and would constitute a nonrepresentative or **biased sample**.

The importance of representative sampling is shown by the failure of a political poll taken during the 1936 presidential election. In that election, Alf Landon was opposing Franklin Roosevelt. The editors of the *Literary Digest* (a now-defunct magazine) conducted a poll by using telephone directories and vehicle registration lists to draw their sample. The final sample consisted of nearly 10 million people! The results showed that Landon would beat Roosevelt by a landslide. Quite to the contrary, Roosevelt soundly defeated Landon. Why was the poll so wrong?

The problem stemmed from the method used to obtain the sample. Fewer people owned a car or telephone in the 1930s than do today. In fact, very few owned either. Those who did own a telephone or car tended to be relatively wealthy and Republican. Consequently, most of the participants polled favored the Republican candidate. Unfortunately for the *Literary Digest,* this sample did not represent the population of voters, and the prediction failed. How could the editors have been so stupid? In fact, they weren't stupid. Such sampling techniques had been used before and worked. It was only in that particular election (in which people were clearly split along party lines) that the problem emerged (Hooke, 1983).

The *Literary Digest* poll failed because it used a biased source (car registration and telephone listings). Whatever source you choose, you should make an effort to determine whether it includes members from all segments of the population in which you have an interest. A good way to overcome the problem of biased source lists is to use multiple lists. For example, you could use the telephone book *and* vehicle registration *and* voter registration lists to select your sample.

Sampling Techniques

At the heart of all sampling techniques is the concept of *random sampling*. In random sampling, every member of the population has an equal chance of appearing in your sample. Whether or not a participant is included in your sample is based on chance alone. Sampling is typically done without replacement. Once an individual is chosen for your sample, he or she cannot be chosen a second time for that sample.

Random sampling eliminates the possibility that the sample is biased by the preferences of the person selecting the sample. In addition, random sampling affords some assurance that the sample does not bias itself. As an example of self-biasing, consider the following case. In 1976, Shere Hite published *The Hite Report: A Nationwide Study on Female Sexuality,* which was a survey of women's sexual attitudes and behaviors. Hite's sample was obtained by initially distributing questionnaires through national mailings to women's groups (the National Organization for Women, abortion rights groups, university women's centers, and others). Later, advertisements were placed in several magazines (the *Village Voice*, *Mademoiselle*, *Brides*, and *Ms.*) informing women where they could write for a copy of the questionnaire. Finally, the questionnaire was reprinted in *Oui* magazine in its entirety (253 women returned the questionnaire from *Oui*).

The question that you should ask yourself at this point is, "Did Hite obtain a random sample of the population of women?" The answer is no. Hite's method had several problems. First, the memberships of the organizations that Hite contacted may not represent the population of women. For example, you cannot assume that members of NOW hold similar views, on the average, to those of the population of all women. Second, asking people through magazine ads to write in for questionnaires further biases the sample. Can you figure out why?

If you said that the people who write in for the questionnaires may be somehow different from those who do not, you are correct. Who would write in to obtain a questionnaire on sexuality? Obviously, women who have an interest in such an issue. In fact, Hite indicates that many of her participants expressed such an interest. One woman wrote, "I answered this questionnaire because I think the time is long overdue for women to speak out about their feelings about sex" (Hite, 1976, p. xxxii). As with the members of the women's organizations, you could question whether the women who wrote in for questionnaires are representative of all women. They probably are not.

When a sample is biased, the data obtained may not indicate the attitudes of the population as a whole. Hite concluded from her sample that women in this country were experiencing a "new sexuality." However, that new sexuality was limited to those women whose attitudes were similar to those who answered her questionnaires.

In 1983, Hite published *The Hite Report on Male Sexuality*. The method she used to gather data was similar to the one used in her earlier study of women. In this book, Hite responded to the criticisms of her method. She presented evidence that her sample of men was similar in age, religion, and education to the most recent census data. What was not clear, however, was whether or not the attitudes of the men who responded to her questionnaire were similar to those of the general population. As in the survey of women, the data obtained may not be representative of the population of men. Some evidence suggests they were not. Hite said that 72% of married men reported having had an extramarital affair. Is this an accurate estimate of the population or an estimate of a special subsection of the population? Apparently, it is the latter. Other surveys have found that about 25% of men report having had extramarital affairs.

The lesson of the Hite example is that you should make every effort to obtain a random sample. This may be difficult, especially if you are dealing with a sensitive topic. You could use some of the strategies previously suggested for reducing nonresponse bias (such as including a small reward or using follow-ups). If your sample turns out to be nonrandom and nonrepresentative, temper any conclusions you draw.

Using the proper sampling technique is one way to obtain a representative sample. Several techniques are available to you. Five of them (simple random sampling, stratified sampling, proportionate sampling, systematic sampling, and cluster sampling) are discussed next. These techniques are not mutually exclusive. Often researchers combine them to help ensure a representative sample of the population.

Simple Random Sampling Randomly selecting a certain number of individuals from the population is a technique called **simple random sampling**. Remember the golf ball example? A simple random sample of 50 would involve dipping your hand into the bag 50 times, each time withdrawing a single ball. Figure 9-5 illustrates the simple random sampling strategy. From the population illustrated at the top of the figure, 10 participants are selected at random for inclusion in your survey.

In practice, selecting a random sample for a survey is more involved than pulling golf balls from a bag. Often it involves consulting a table of random numbers. The numbers in such a table have been chosen at random and then subjected to a number of statistical tests to ensure that they have the expected properties of random numbers. You can find a table of random numbers in the Appendix (Table 1A).

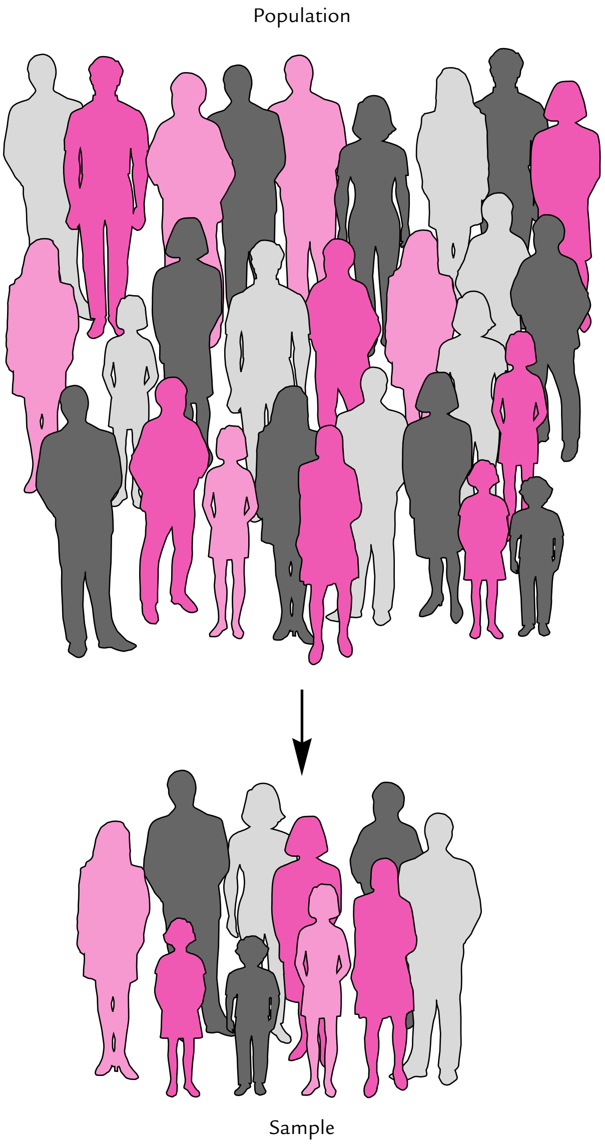


FIGURE 9-5 Example of simple random sampling. The people at the top of the figure represent the population, and the people at the bottom represent the randomly selected sample.

As an example of how to use the table of random numbers to select a random sample, imagine you are using the telephone book as a source list. Starting on any page of the random number table, close your eyes and drop your finger on the page. Open your eyes and read the number under your finger. Assume that the number is 235,035. Then go to page 235 in the telephone book and select the 35th name on that page. Repeat this process until you select all the participants constituting the sample.

A variant of random sampling that can be used when conducting a telephone survey is *random digit dialing* (Dillman, 2000). List all the exchanges in a particular area (the first three digits of the phone numbers, not including the area code). You then use the table of random numbers or a computer to select four-digit numbers (e.g., 5,891). The exchange plus the four-digit number provides the number to be called. (Any nonworking numbers are discarded.) This technique allows you to reach unlisted as well as listed numbers.

Even though random sampling reduces the possibility of systematic bias in your sample, it does not guarantee a representative sample. You could, quite at random, select participants who represent only a small segment of the population. In the golf ball example, you might select 50 orange golf balls. White and yellow golf balls, even though represented in the population, are not in your sample. One way to combat this problem is to select a large sample (such as 200 rather than just 50 balls). A large sample is more likely to represent all segments of the population than a small one. However, it does not *guarantee* that representation in your sample will be proportionate to representation in the population. You may end up with 90 white, 90 orange, and only 20 yellow golf balls in a sample of 200, although such a result is highly unlikely. In addition, as you increase sample size, you also increase the cost and time needed to complete the survey. Fortunately, more sophisticated techniques provide a random yet representative sample without requiring a large number of participants.

QUESTIONS TO PONDER

1. What is a representative sample and why is it important to have one for a survey?
2. What is a biased sample and how can a biased sample affect your results?
3. What is a random sample and why is it important to do random sampling?
4. What is simple random sampling?

Stratified Sampling **Stratified sampling** provides one way to obtain a representative sample. You begin by dividing the population into segments, or *strata* (Kish, 1965). For example, you could divide the population of a particular town into Whites, Blacks, and Hispanics. Next, you select a separate random sample of equal size from each stratum. Because individuals are selected from each stratum, you guarantee that each segment of the population is represented in your sample. Figure 9-6 shows the stratified sampling strategy. Notice that the population has been divided into two segments (gray and colored figures). A random sample is then selected from each segment.

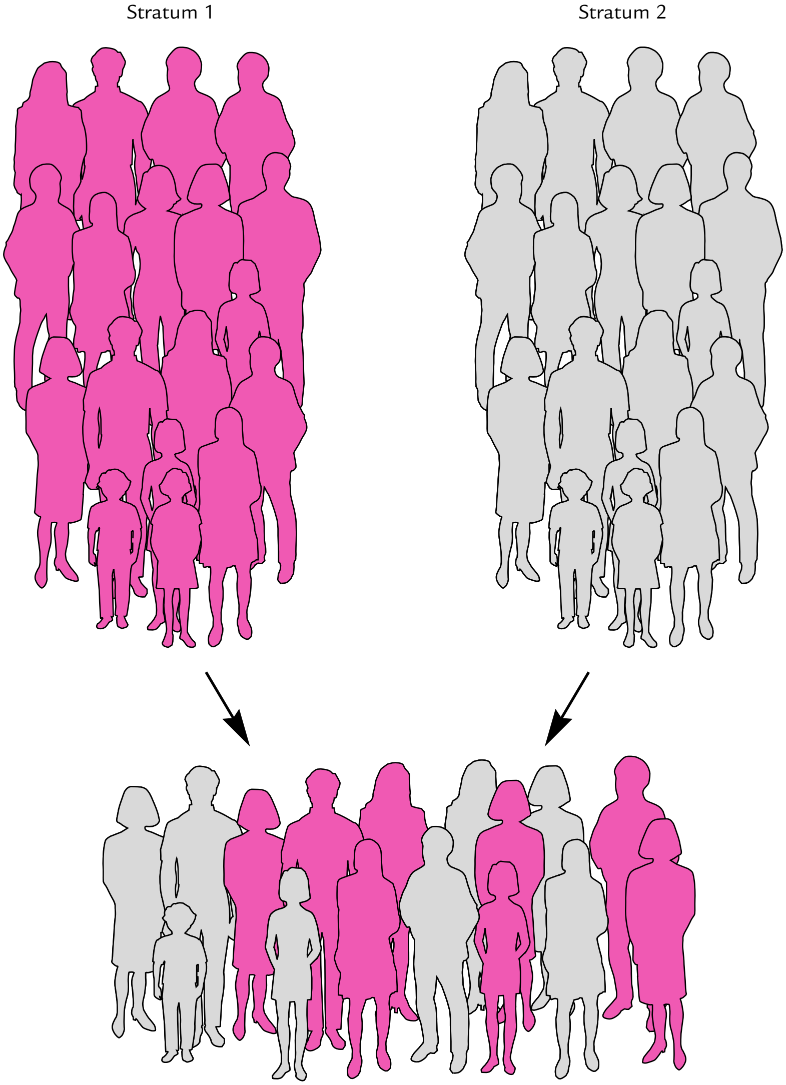


FIGURE 9-6 Example of stratified sampling. The population is divided into two strata from which independent random samples are drawn.

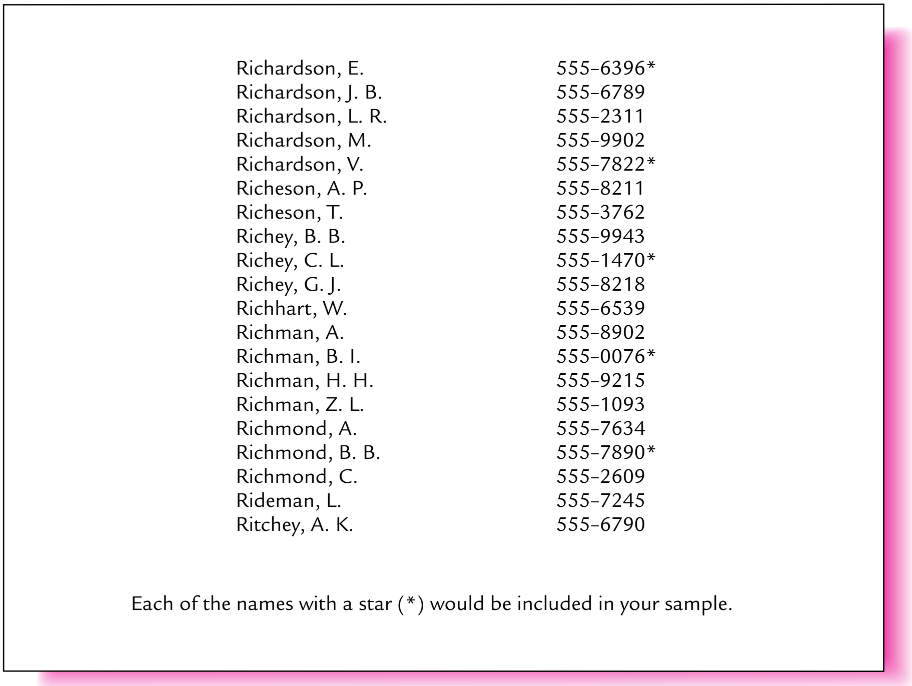
Proportionate Sampling Simple stratified sampling ensures a degree of representativeness, but it may lead to a segment of the population being overrepresented in your sample. For example, consider a community of 5,000 that has 500 Hispanics, 1,500 Blacks, and 3,000 Whites. If you used a simple stratification technique in which you randomly selected 400 people from each stratum, Hispanics would be overrepresented in your sample relative to Blacks and Whites, and Blacks would be overrepresented relative to Whites. You could avoid this problem by using a variant of simple stratified sampling called **proportionate sampling**.

In proportionate sampling, the proportions of people in the population are reflected in your sample. In the population example, your sample would consist of 10% Hispanics (500/5,000 = 10%), 30% Blacks (1,500/5,000 = 30%), and 60% Whites (3,000/5,000 = 60%). So, if you draw a sample of 1,200, you would have 120 Hispanics, 360 Blacks, and 720 Whites. According to Kish (1965), this technique is the most popular method of sampling.

By the way, stratification and proportionate sampling can be done after a sample has been obtained (Kish, 1965). You randomly select from the participants who responded the number from each stratum needed to match the characteristics of the population.

Systematic Sampling **Systematic sampling** is a popular technique that is often used in conjunction with stratified sampling (Kish, 1965). Figure 9-7 illustrates the systematic sampling technique.

According to Kish (1965), this technique involves sampling every *k*th element after a random start. For example, once you have randomly chosen the page of the telephone book from which you are going to sample, you then might pick every fourth item (where $k = 4$). Systematic sampling is much less time consuming and more cost effective than simple random sampling. For example, it is much easier to select every fourth item from a page than to select randomly from an entire list.



Richardson, E.	555-6396*
Richardson, J. B.	555-6789
Richardson, L. R.	555-2311
Richardson, M.	555-9902
Richardson, V.	555-7822*
Richeson, A. P.	555-8211
Richeson, T.	555-3762
Richey, B. B.	555-9943
Richey, C. L.	555-1470*
Richey, G. J.	555-8218
Richhart, W.	555-6539
Richman, A.	555-8902
Richman, B. I.	555-0076*
Richman, H. H.	555-9215
Richman, Z. L.	555-1093
Richmond, A.	555-7634
Richmond, B. B.	555-7890*
Richmond, C.	555-2609
Rideman, L.	555-7245
Ritchey, A. K.	555-6790

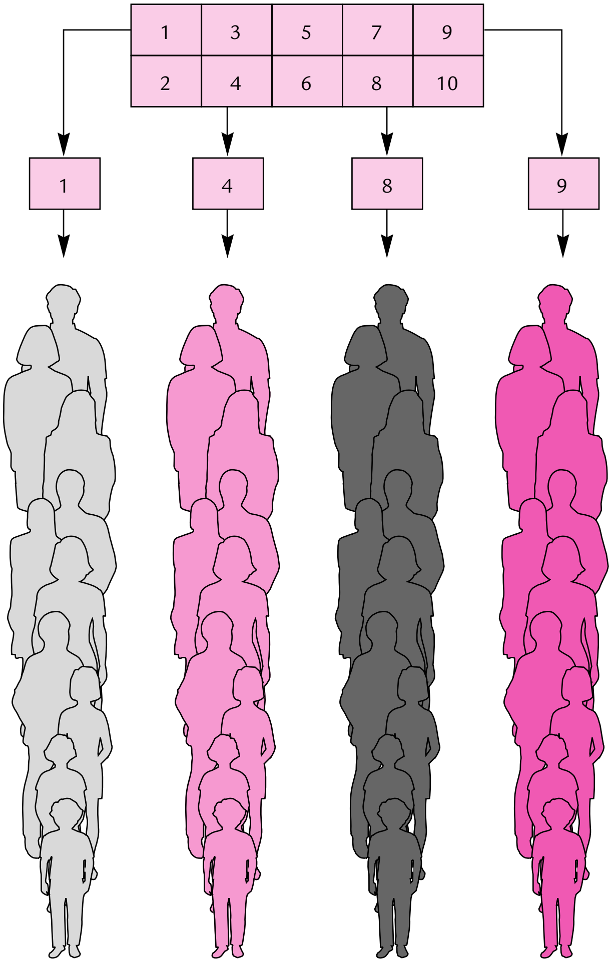
Each of the names with a star (*) would be included in your sample.

FIGURE 9-7 Example of systemic sampling. After a random start, every selected name is included in the sample (indicated with an asterisk).

Cluster Sampling In some cases, populations may be too large to allow cost-effective random sampling or even systematic sampling. You might be interested in surveying children in a large school district. To make sampling more manageable, you could identify naturally occurring groups of participants (clusters) and randomly select certain clusters. For example, you could randomly select certain departments or classes from which to sample. Once the clusters have been selected, you would then survey all participants within the clusters. **Cluster sampling** differs from the other forms of sampling already discussed in that the basic sampling unit is a group of participants (the cluster) rather than the individual participant (Kish, 1965). Figure 9-8 illustrates cluster sampling. This figure shows how you select four groups from a larger pool of groups.

An obvious advantage to cluster sampling is that it saves time. It is not always feasible to select random samples that focus on single elements (individuals, families, etc.). Cluster sampling provides an acceptable, cost-effective method of acquiring a

FIGURE 9-8 Example of cluster sampling. After selecting subgroups of the population, all participants in each subgroup are surveyed.



sample. On the negative side, cluster sampling does limit your sample to those participants found in the chosen clusters. If participants within clusters are fairly similar to one another but differ from those in other clusters, the sample will leave out important elements of the population. For example, clusters consisting of geographical areas of the United States (e.g., East, Midwest, South, Southwest, and West) may differ widely in political opinion. If only East and Midwest are selected for the sample, the opinions collected may not reflect the opinions of the country as a whole. Thus, cluster sampling does have drawbacks.

A variant of cluster sampling is **multistage sampling**. You begin by identifying large clusters and randomly selecting from among them (first stage). From the selected clusters, you then randomly select individual elements (rather than selecting all elements in the cluster). This method can be combined with stratification procedures to ensure a representative sample.

Other sophisticated sampling techniques are available to the survey researcher, but to explore them all would require a whole book. If you are interested in learning about these techniques, read Kish (1965).

Random and Nonrandom Sampling Revisited

In Chapter 6, we distinguished between random sampling (in which each member of a population has an equal chance of being selected) and nonrandom sampling (in which a limited group of potential participants is tapped). The sampling techniques we have just discussed may be used in the context of random or nonrandom sampling. Ideally, you would want to use random sampling. This is especially true, as noted in Chapter 6, if you want to make specific predictions about specific behaviors. However, as a practical matter, it may not always be possible to use a true random sample. Instead, you may have to administer your questionnaire to a convenience sample, such as students at a particular university, which is a nonrandom sample. Similarly, surveys conducted via the Internet use nonrandom samples, consisting only of those with computers who know how to access the Internet and have the ability to complete the survey. Of course, using a nonrandom sample limits the generality of your results, and making specific predictions about behavior may not be possible. However, a nonrandom sample (as noted in Chapter 6) is perfectly acceptable for most research interests in psychology. If you use nonrandom sampling, you should include a discussion of possible limitations of your results in the discussion section of any report that you write.

QUESTIONS TO PONDER

1. What are the various sampling techniques that represent modifications of simple random sampling?
2. Under what conditions would you use each of the sampling techniques discussed above?
3. What are the implications of using a nonrandom sample?

Sample Size

One factor you must contend with if you perform a survey is the size of your sample. You should try to select an *economic sample*—one that includes enough participants to ensure a valid survey and no more. You must take into account two factors when considering the size of the sample needed to ensure a valid survey: the amount of acceptable error and the expected magnitude of the population proportions.

The question of acceptable error arises because most samples deviate to some degree from the population. If you conduct a political poll on a sample of 1,500 registered voters and find that 62% of the sample favors Smith and 38% Jones, you would like to say that 62% of the population favors Smith. However, these sample proportions do not exactly match those of the population (the population proportions may be 59% and 41%). This deviation of sample characteristics from those of the population is called **sampling error**.

When determining sample size, you must decide the acceptable amount of sampling error. Unfortunately, there are no broad rules of thumb as to the acceptable margin of error. It depends in part on the use to which you will put your results (Moser & Kalton, 1972). If you plan to apply your results to implement changes in behavior, you may want a small margin of error. If you are interested simply in describing a set of characteristics, you may tolerate a larger margin of error. A good way to determine the acceptable margin of error is to look at literature describing similar surveys to see what margin of error was used.

The second component you need to consider when determining sample size is the magnitude of the differences you expect to find. Here again, there is no broad rule of thumb to guide you. You can make use of previous surveys to get an estimate of the magnitude of the differences. Or you can conduct a small pilot survey to gain some insight into the magnitudes.

Once you have determined the acceptable error and the expected magnitude of differences, you can calculate the size of the sample needed. The calculation is relatively easy for simple random sampling. Moser and Kalton (1972) suggested the following formula:

$$n' = \frac{P'(1 - P')}{(SE_p)^2}$$

where P is the estimate of the proportion of the population that has a particular characteristic and SE_p is the acceptable margin of error. For example, if you expect 62% of the population to favor Smith in an election and your acceptable margin of error is 2% (0.02), then the formula gives $n = 589$. Thus, you should have 589 participants in your sample.

When the size of the population is large, you do not need to consider population size when calculating sample size. If the population is small, however, then you must use the *finite population correction* (fpc) when calculating sample size. Crano and Brewer (1986) suggest using the following formula when the sample size is more than 10% of the population size:

$$n = N \times n'/(N + n')$$

where n = the corrected sample size, n' = the sample size calculated with the previous formula, and N = the size of the population from which the sample is to be drawn. For example, using the previous numbers and N = 2,000, you have

$$n = 2000 \times 589/(2000 + 589) = 455$$

Thus, if the population from which your sample will be drawn consists of only 2,000 participants, you would use a sample size of 455 rather than 589.

For stratified sampling, determining sample size is more difficult than for simple random sampling. You must take into account the between-strata error (the variability in the scores of participants in different strata) and the within-strata error (the variability in the scores of participants within the same stratum). The formulas for computing sample size with the more sophisticated sampling techniques are complex. If you pursue survey research using these techniques, consult Moser and Kalton (1972) and Kish (1965) for more information.

QUESTIONS TO PONDER

1. What is meant by an "economic sample"?
2. What is sampling error and how do you know if you have an acceptable level?
3. How does the magnitude of the differences you expect to observe affect your decision about sample size?
4. What are some of the sample size issues you need to consider for different sampling techniques?

SUMMARY

Survey research is used to evaluate the behavior (past, present, and future) and attitudes of your participants. Survey research falls into the category of correlational research. Therefore, you cannot draw causal inferences about behavior from your survey data, no matter how compelling the data look. Surveys are used in a wide variety of situations. They can be used to research the marketability of a new product, to predict voter behavior, or to measure existing attitudes on a variety of issues.

The first step in a survey is to clearly define the goals of your research. Your questionnaire is then designed around those goals. You should have a reasonably focused goal for your survey. A questionnaire that tries to do too much may be confusing and burdensome to your participants. Keep your questionnaire focused on the central issues of your research.

Often a questionnaire is organized so that questions about your participants' characteristics (demographic items) and questions about the behavior or attitude of interest are included. The demographic items can later be used as predictor variables when you look for relationships among the variables that you measured.

Questionnaire items can be of several types. Open-ended questions allow your participants to answer in their own words. A major advantage of this type of question is the richness of the information obtained. A drawback is that responses are difficult to summarize and analyze. A restricted question provides response categories for participants. A variation on the restricted item is a rating scale on which participants circle a number reflecting how they feel. This type of item yields data that are easier to summarize and analyze. However, the responses made to restricted items are not as rich as those obtained with an open-ended item. A partially open-ended item gives participants not only clearly defined response alternatives but also a space to write in their own response category.

Once you have decided what types of items to include on your questionnaire, you must then actually write your questions. When writing items, you should avoid using overly complex words when simpler words will suffice. Your questions should be precise. Vague or overly precise wording yields inconsistent data. In addition, you should avoid using words that are biased or judgmental.

A questionnaire is more than just a collection of questions. Questions should be presented in a logical order so that your questionnaire has continuity. Also, it is a good idea to place demographic items at the end. These questions tend to be boring, and participants may be turned off if you have demographic items at the beginning of your questionnaire. Sensitive questions should be placed toward the middle. Your participants may be more willing to answer such questions after answering several other, more innocuous questions. Sensitive items should be carefully worded. Your questionnaire should have a logical "navigational path." This path should lead the respondent through the questionnaire as if he or she were reading a book.

Constructing a questionnaire involves more than sitting down and writing a set of items. Developing a good questionnaire involves several steps, including assessing its reliability, or your questionnaire's ability to produce consistent results. One way to assess reliability is to administer your questionnaire (or parallel forms of the questionnaire) more than once. If the results are highly similar, the questionnaire is reliable. Another way to assess reliability is with a single administration of your questionnaire. The most common way to do this is to use a split-half method by which you divide your questionnaire in half (e.g., odd versus even items) and correlate the two halves. Two statistics used to evaluate split-half reliability are the Kuder–Richardson formula and coefficient alpha.

If you find low reliability, you can do several things to increase it. You can increase the number of items on your questionnaire, standardize administration procedures, make sure that you score questions carefully, and ensure that your items are clear, well written, and appropriate for your sample.

In addition to assessing reliability, you should evaluate the validity of your questionnaire. The term *validity* in this context refers to whether your questionnaire actually measures what you intend it to measure. There are three ways to assess validity. First, you can establish content validity by making sure that items on your questionnaire cover the full range of issues relevant to the phenomenon you are studying. Second, criterion-related validity can be established by correlating the results from your questionnaire with one of established validity. Third, you can establish construct validity by establishing that the results from your questionnaire match well with predictions made by a theory. No one of these methods is best. Perhaps the best approach is to establish validity using more than one of the three methods.

Five ways to administer your questionnaire are the mail survey, group administration, telephone survey, face-to-face interview and Internet survey. The mail survey is easiest. You simply mail your questionnaires and wait for a response. However, this method is plagued by nonresponse bias. Return rates can be increased with effective cover letters, follow-up reminders tailored to the nature of your participant population, and small rewards. In group administration, you give your questionnaire to a large number of participants at once. The advantage of group administration is that you can collect large amounts of data quickly. Surveys also can be conducted over the telephone. Questionnaires designed for telephone surveys should be relatively short, with clearly worded, short questions. Because your questions will be read to your participants, make sure that the person reading the questions speaks clearly and slowly. In an interview, you ask your questions to your participants in a face-to-face session. Interviews can be either structured (questions asked from a prepared questionnaire in a fixed order) or unstructured (each interview is different). Finally, you can conduct your survey on the Internet, which allows you to reach large numbers of potential respondents. Data can be collected quickly and easily via the Internet. However, the sample obtained from the Internet may not be representative, and you must have the equipment, resources, and knowledge necessary to post a questionnaire this way.

One of the most crucial stages of survey research is acquiring a sample of participants. Because you want to make statements about how people think on an issue, be sure your sample represents the population. Biased samples lead to invalid data and ultimately incorrect conclusions. Sampling techniques include simple random sampling (in which every participant has an equal chance of being in your survey) and stratified sampling (in which your population is broken into smaller segments and random samples are then drawn from those smaller segments). Other sampling techniques are proportionate sampling, multistage sampling, and cluster sampling. The sampling technique you use depends on the needs of your survey.

Whichever sampling technique you choose, you must consider the issue of sample size. Your sample should be large enough to be representative of the population, yet not too large. Try to acquire an economic sample that has just enough participants to adequately assess behavior or attitudes. The size of the most economic sample is determined with a special formula.

KEY TERMS

open-ended item
restricted item
partially open-ended item
mail survey
nonresponse bias
Internet survey
telephone survey
face-to-face interview
representative sample
biased sample
simple random sampling
stratified sampling
proportionate sampling
systematic sampling
cluster sampling
multistage sampling
sampling error

10

CHAPTER

Using Between-Subjects and Within-Subjects Experimental Designs

CHAPTER OUTLINE

- Types of Experimental Design
- The Problem of Error Variance in Between-Subjects and Within-Subjects Designs
  - Sources of Error Variance
  - Handling Error Variance
- Between-Subjects Designs
  - The Single-Factor Randomized-Groups Design
  - Matched-Groups Designs
- Within-Subjects Designs
  - An Example of a Within-Subjects Design: Does Caffeine Keep Us Going?
  - Advantages and Disadvantages of the Within-Subjects Design
  - Sources of Carryover
  - Dealing With Carryover Effects
  - When to Use a Within-Subjects Design
  - Within-Subjects Versus Matched-Groups Designs
  - Types of Within-Subjects Designs
- Factorial Designs: Designs With Two or More Independent Variables
  - An Example of a Factorial Design: Can That Witness Really *Not* Remember an Important Event?
  - Main Effects and Interactions
  - Factorial Within-Subjects Designs
  - Higher-Order Factorial Designs
- Other Group-Based Designs
- Designs With Two or More Dependent Variables
- Confounding and Experimental Design
- Summary
- Key Terms

As we pointed out in Chapters 1 and 3, a major goal of research is to establish clear causal relationships between variables. The correlational research designs discussed in Chapters 8 and 9 identify potential causal relationships and often are used when causal variables cannot or should not be manipulated directly. However, correlational designs are simply not adequate for establishing causal relationships between variables.

When your goal is to establish causal relationships and you can manipulate variables, an experimental research design is used. By manipulating an independent variable while rigidly controlling extraneous factors, you can determine whether this manipulation causes changes in the value of the dependent variable.

TYPES OF EXPERIMENTAL DESIGN

In Chapter 4, we noted that every true experiment contains an independent variable (also referred to as a *factor* in experimental terminology), which you manipulate, and a dependent variable, which you observe and measure. To manipulate the independent variable, you set its value to at least two different values or "levels" during the course of the experiment and observe your subjects' performances under each level. You then compare these performances. If you can show that performance differed across the levels of the independent variable and that these differences are reliable, you can conclude that a change in the level of the independent variable *causes* a change in the value of the dependent variable.

There are two ways in which you can manipulate your independent variable. You can vary it quantitatively by changing the amount of the variable to which each group of participants is exposed. For example, in an experiment testing the effect of different doses of Prozac on memory, you could vary the amount of Prozac administered to your participants by giving doses of 10 milligrams (mg), 20 mg, and

30 mg. You also can vary your independent variable qualitatively. For example, in an experiment testing the effects of different antidepressants on memory, you could give participants in your different treatment groups Prozac, Lexapro, or Zoloft.

The simple logic of manipulating an independent variable and observing related changes in behavior is at the heart of every experimental design. However, to deal with the complexities of real-world research problems, researchers have developed a wide variety of experimental designs. We can simplify the situation somewhat by noting that experimental designs can be categorized into three basic types: between-subjects, within-subjects, and single-subject designs. In a **between-subjects design**, different groups of subjects are randomly assigned to the levels of your independent variable. In a **within-subjects design**, a single group of subjects is exposed to *all* levels of your independent variable. In both the between-subjects and within-subjects designs, data from subjects within a given treatment are averaged and analyzed. A **single-subject design** is similar to the within-subjects design in that subjects are exposed to all levels of the independent variable. The main difference from the within-subjects design is that you do not average data across subjects. Instead, you focus on changes in the behavior of a single subject (or a small number of individual subjects) under the different treatment conditions.

In this chapter, we discuss between-subjects and within-subjects designs. (Single-subject designs are discussed in Chapter 12.) The plan of this chapter is to discuss, first, the problem of error variance in experimental design and how it is handled. We then introduce single-factor between-subjects and within-subjects designs, designs that include only one independent variable. Finally, we explore between-subjects and within-subjects designs that include two or more independent variables.

THE PROBLEM OF ERROR VARIANCE IN BETWEEN-SUBJECTS AND WITHIN-SUBJECTS DESIGNS

Error variance is the variability among scores caused by variables other than your independent variables (extraneous variables or subject-related variables such as age, gender, and personality). The problems posed by error variance are common to all three experimental designs. However, each design has its own way of dealing with error variance. In this chapter, we focus on how we deal with error variance in between-subjects and within-subjects designs. In Chapter 12, we discuss how error variance is handled in single-subject designs.

Sources of Error Variance

In the real world, it is rarely possible to hold constant all the extraneous variables that could affect the value of your dependent variable. Subjects in your experiment differ from one another in innumerable ways that could individually or collectively affect their scores on the dependent measure, the environmental conditions are not absolutely constant, and even the same subject will not be exactly the same from moment to moment. To the extent that these variations affect your dependent variable, they induce fluctuations in scores that have nothing to do with your independent variable. That is, they produce error variance.

TABLE 10-1 Scores from Hypothetical THC Experiment

	PERFORMANCE ON DEPENDENT MEASURE	
	Control Group	*Experimental Group*
	25	13
	24	19
	18	22
	29	18
	19	23
Mean	23	19

An example may help clarify this concept. In an experiment on the effects of THC (the active ingredient in marijuana) on a simulated air traffic–control task, one group is exposed to a dose of THC (the experimental group) and one is not (the control group). Within each group, all participants would have been exposed to the same level of the independent variable. Yet it is unlikely that all participants in a group would turn in the same scores on the dependent measure (score on the simulated air traffic–controller task). Participants differ from one another in many ways that affect their performance. Some may be more resistant to THC, have better attention skills, or have greater perceptual abilities than others, for example. The variation in scores produced by these uncontrolled variables is the error variance that we are discussing.

Table 10-1 shows the scores turned in by participants in this hypothetical experiment. The scores for each group have been averaged, and the means are presented at the bottom of the table. Judging from the means, it appears that THC reduced the participants' scores on the dependent variable. However, given the variability in scores evident within each group, it seems plausible to suggest that the difference in the means may reflect nothing more than preexisting participant differences that did not quite balance out across the two conditions of the experiment. The problem is that you cannot tell, simply by looking at the means, which explanation is correct. The problem of error variance is therefore serious. It affects your ability to determine the effectiveness of your independent variable.

QUESTIONS TO PONDER

1. How do between-subjects, within-subjects, and single-subject experiments differ?
2. What are the sources of error variance in a between-subjects design and how might error variance affect your results?

Handling Error Variance

Fortunately, there are ways you can cope with the problem of error variance. You can take steps to reduce error variance, you can take steps to increase the effect of your independent variable, and you can randomize error variance across groups. Let's look at each of these strategies in more detail.

Reducing Error Variance The principal way to reduce error variance is to hold extraneous variables constant by treating subjects within a group as similarly as possible. For example, you could test participants in an isolated room to eliminate outside distractions and make sure that you read instructions to all participants within a group in the same way. You should also follow the same procedures for all subjects within a group. Error variance also can be reduced by using subjects matched on characteristics that you believe contribute to error variance. For example, you could use participants who are of the same age or educational level. Although this may reduce external validity, you can always relax the restrictions in a later experiment. The first priority is to obtain reliable results. A similar tactic is to match subjects across groups on some characteristic relating to the dependent variable in a matched-groups design or use the same subjects for all levels of your independent variable in a within-subjects design. (We discuss matching and using within-subjects designs later in this chapter.)

Increasing the Effectiveness of Your Independent Variable Another way to deal with error variance is to select the correct levels of your independent variable for your experiment. A weak manipulation (e.g., too low a dose of THC) may not influence your dependent variable, leaving the effect of your independent variable buried in whatever amount of error variance exists. Of course, it is difficult to know beforehand just how to manipulate your independent variable. You can get some idea about the levels to include from previous research and by conducting a pilot study before you run your actual experiment. You also might consider using a dependent variable that is sensitive enough to detect the effects of your independent variable.

Randomizing Error Variance Across Groups Regardless of the steps that you take to minimize error variance, you can never eliminate it completely. In between-subjects designs, you can reduce any remaining error variance by randomizing error variance across groups. This is accomplished through random assignment of subjects to your treatment conditions. As noted in Chapter 4, *random assignment* means that subjects are assigned to groups on a random basis so that each subject has an equal chance of appearing in any group in your experiment. You could do this by drawing participants' names out of a hat and assigning the first name pulled to the Experimental Group, the second name to the Control Group, and so on. In an actual experiment, you would probably accomplish random assignment by using a table of random numbers rather than by drawing names out of a hat. In either case, random assignment results in groups of subjects that have been equalized, over the long run, on individual difference factors (e.g., intelligence and gender), resulting in error variance being evenly distributed across groups.

Statistical Analysis Although random assignment *tends* to equalize error variance across groups, there is no guarantee that it *will* do so. Similarly, despite your best efforts to eliminate error variance, some will remain. How, then, can you determine whether an effect observed in your data was caused by your manipulation and not by error variance? Although you can never be sure, you can estimate the *probability* with which error variance alone would produce differences between groups at least as large as those actually observed. You do this by subjecting your data to a statistical analysis using *inferential statistics* (see Chapter 14). If this probability is low enough, your results are said to be *statistically significant*, and you conclude that your results were most likely due to the manipulation of your independent variable and not error variance.

QUESTIONS TO PONDER

1. What steps can you take to deal with error variance in a between-subjects design?
2. How are statistics used to test the reliability of data from a between-subjects experiment?

BETWEEN-SUBJECTS DESIGNS

The time has come to examine the types of between-subjects designs available to you. We begin with single-factor designs, in which you manipulate only one independent variable.

The Single-Factor Randomized-Groups Design

A commonly used form of the between-subjects design is the *randomized-groups design*. When using this design, you randomly assign subjects to the levels of your independent variable to form "groups" of subjects. There are two variants of the randomized-groups design: the randomized two-group design and the randomized-multigroup design. We explore these designs next.

The Randomized Two-Group Design If you randomly assign your subjects to two groups, expose the two groups to different levels of the independent variable, and take steps to hold extraneous variables constant, you are using a **randomized two-group design**. Figure 10-1 illustrates the basic steps to follow when conducting a randomized two-group experiment. Begin by sampling a group of subjects from the general population (top). Then, randomly assign the participants from this group into your two treatment groups. Next, expose the participants in each group to their treatments and record their responses. Compare the two means to determine whether they differ. Finally, submit the data to a statistical analysis to assess the reliability of any difference that you find.

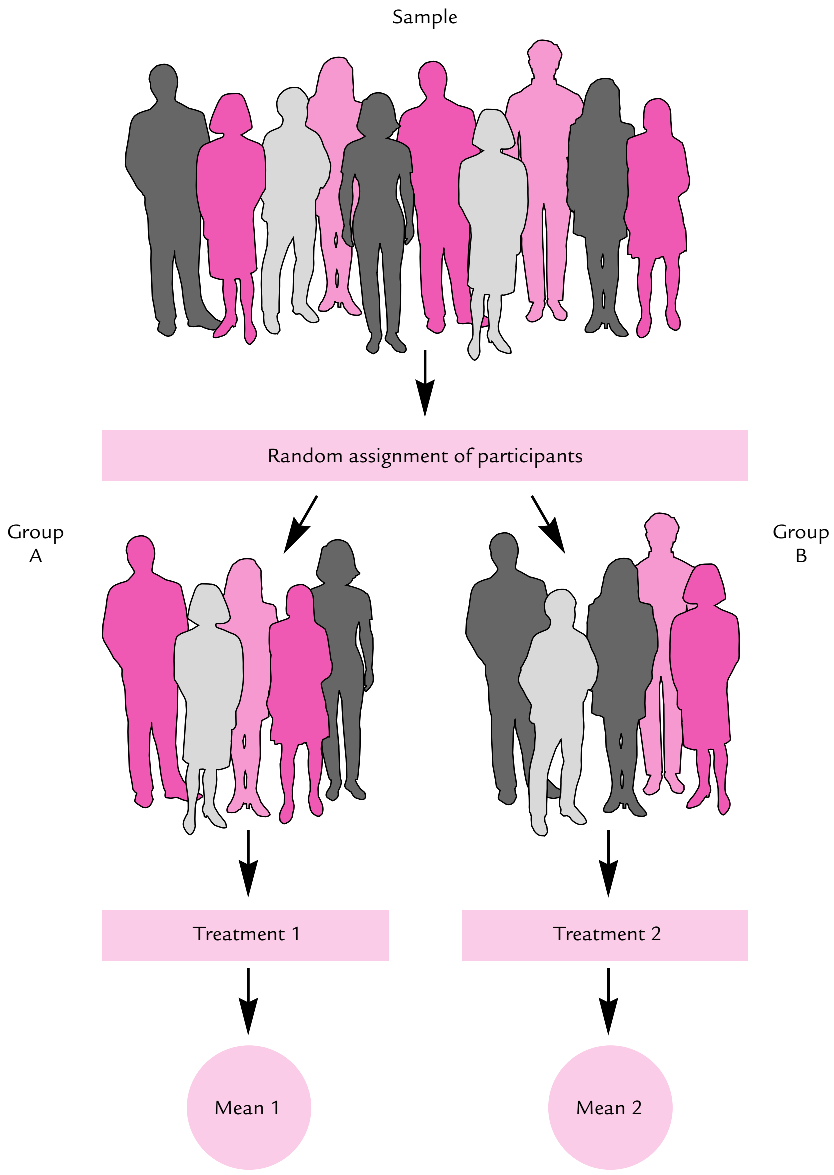


FIGURE 10-1 A completely randomized two-group experimental design.

An experiment conducted by Jo-Ann Tsang (2006) provides an excellent example of an experiment using a randomized two-group design. Tsang was interested in investigating whether gratitude resulted in more prosocial behavior than mere positive emotion. Participants in Tsang's experiment were told that they would be playing a game in which they would be allocating resources to another participant in the study (in reality, there was no other participant). Participants were told further that the game would be played in three rounds. Resource allocations were made by writing down an amount of money to allocate to the fictitious participant on a slip of paper, which would be taken to the fictitious participant by the experimenter.

Tsang (2006) randomly assigned the real participants to one of two conditions. In the "favor condition," participants were told that their partner had allocated $9 of $10 to them and kept $1 for himself in the second round. They were also given a note saying, "I saw that you did not get a lot in the last round—that must have been a bummer" (Tsang, 2006, p. 142). In the "chance control condition," participants were told that they had received the $9 by chance and that their partner had received $1 by chance. No note accompanied the distribution information. The measure of prosocial behavior directed at the fictitious participant was the amount of money that the real participant allocated (out of $10) in Round 3. Tsang found that participants allocated significantly more money to the fictitious other participant in the favor condition (M = $7.38) than in the chance control condition (M = $5.84).

The randomized two-group design is one of the simplest available, yet it has several advantages over other, more complex designs. First, it is simple to carry out. As was the case in Tsang's (2006) experiment, you need only two levels of your independent variable. Second, everything else being equal, it requires relatively few subjects. For example, Tsang used only 40 participants in her experiment. An experiment with these few subjects is relatively economical in terms of time and materials. Third, no pretesting or categorization of subjects is necessary. The randomized-group strategy often is more than adequate to test your hypothesis, avoiding the need for a more complex matching strategy (see the section on matched-groups designs later in this chapter). Finally, statistical analysis of the resulting data is relatively simple. Indeed, some electronic calculators have the required statistics built into them, so you need only enter the data and press the appropriate button.

A disadvantage of the randomized two-group design is that it provides a limited amount of information about the effect of the independent variable. You learn only a few things, such as whether the two groups differed (on the average) in their responses to the independent variable under the two levels tested, in what direction, and by how much. For example, based on the results of Tsang's (2006) experiment, all you know is that believing that someone gave you $9 of $10 increased prosocial behavior. But, how would other allocations (e.g., $6 out of $10) affect prosocial behavior? You do not learn much about the nature of the relationship, or *function*, relating the independent and dependent variables.

This point can be illustrated with an experiment by Gold (1987). Gold was interested in determining whether glucose (blood sugar) affects memory. In Gold's experiment, rats were individually placed on the white side of a rectangular box that was divided into a well-lit white compartment and a dimly lit black compartment. Because rats tend to prefer darkness over light, they quickly crossed into the black compartment, where they received a mild foot shock. Immediately after this experience, the rats were each injected with glucose. Different groups received different amounts of the glucose. The rats were then returned to their home cages. Twenty-four hours later, the animals were again placed in the white compartment and the amount of time they took to reenter the black compartment was recorded. The rats should have been hesitant to reenter to the extent that they remembered the shock they had

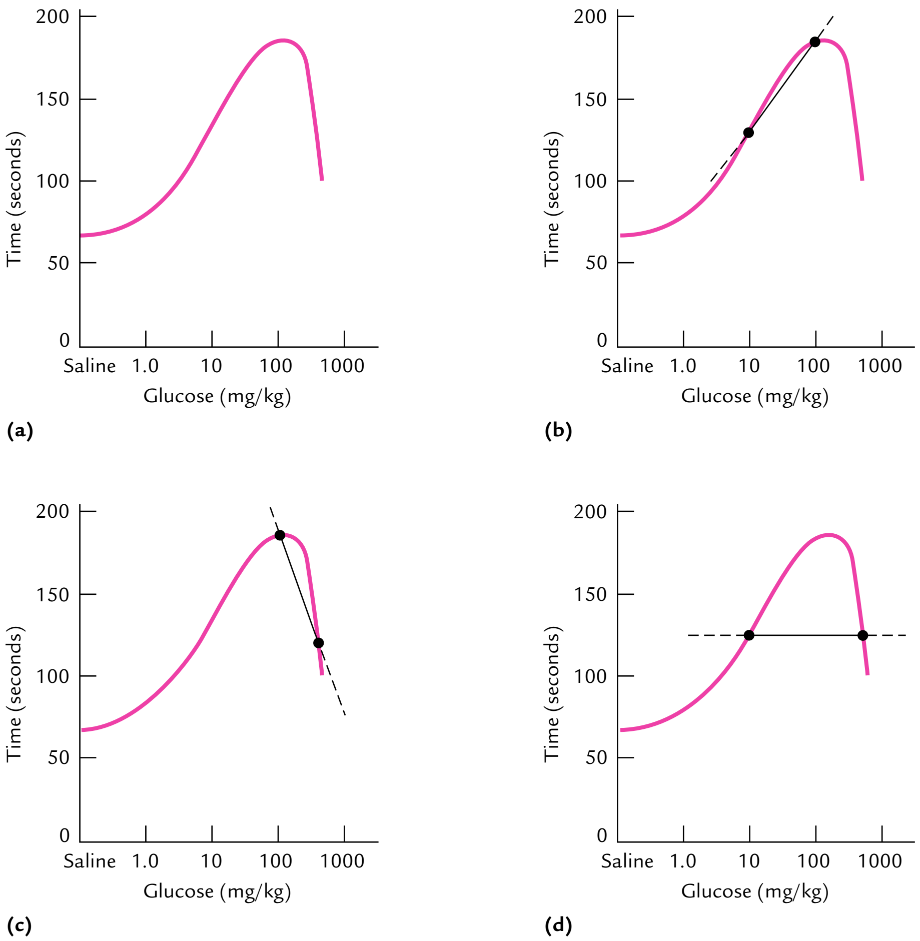


FIGURE 10-2 (a) Results of experiment relating glucose dosage to memory (as measured by the time required to enter a dark compartment); (b, c, and d) three functions based on Gold's data, showing lines estimated from various pairs of points.
SOURCE: Panel (a) from Gold, 1987; reprinted with permission.

received on the previous day. Thus, greater amounts of delay to reenter should have reflected better memory for the shock.

Figure 10-2 shows, in idealized form, the results of Gold's (1987) experiment. In panel (a), the mean number of seconds to reenter the black compartment is plotted against glucose dose. Glucose did affect memory and in a dose-dependent manner. The function relating glucose dose to reentry time is shaped somewhat like an inverted U, with intermediate doses being more effective than higher or lower doses. Gold concluded that glucose can be used in some cases to improve memory (if it's not overdone).

Although Gold's (1987) experiment used several groups, imagine that Gold had used only two. Panel (b) shows what Gold's results would have looked like had he chosen to use glucose doses of 10 milligrams/kilogram (mg/kg) and 100 mg/kg of body weight. What would Gold have concluded?

Panel (c) shows what Gold's (1987) results would have looked like had he chosen to use 100 mg/kg and 600 mg/kg doses. What would Gold's conclusion have been in this case?

Finally, panel (d) shows Gold's (1987) results had he chosen 10 mg/kg and 600 mg/kg doses. What would the conclusion have been now?

If you were unaware of the inverted U-shaped function relating memory to glucose level, it might seem that these three experiments had yielded contradictory results. Furthermore, if you attempted to extrapolate the function beyond the two data points collected in a given experiment—dashed lines in panels (b), (c), and (d)—you would form an erroneous picture of the relationship.

This problem can be solved by conducting a series of two-group experiments in which different levels of the independent variable are chosen for each experiment. However, more efficient designs for sweeping out a functional relationship are available and will be examined later.

A second limitation of the randomized two-group design concerns its sensitivity to the effect (if any) of the independent variable. In cases in which subjects differ greatly from one another on characteristics that influence their performances on the dependent measure, these variations may make it difficult to detect the effect of the independent variable. In such cases, the randomized two-group design may indicate no effect of the independent variable although one was actually present. (The solution is to use a matched-pairs design, which we describe later in the chapter.)

Finally, when you are interested in investigating the limits of an effect, two groups are rarely enough. You must include several levels of an independent variable to adequately test the more subtle effects of your independent variable.

The Randomized-Multigroup Design One way to expand the randomized two-group design is to add one or more levels of the independent variable. You can of course include as many levels of your independent variable as needed to test your hypothesis. As we noted earlier, there are two ways to manipulate your independent variable: quantitatively or qualitatively. When you manipulate your independent variable quantitatively, you are using a **parametric design**. The term *parametric* refers to the systematic variation of the amount of the independent variable. (This use of the term must be distinguished from the use of the word *parametric* to denote a class of inferential statistics.) Manipulating your independent variable qualitatively results in a **nonparametric design**.

A variation on the single-factor multigroup design is one that includes multiple control groups. This design is used when a single control group is not adequate to rule out alternative explanations of your results, and it is known as the **multiple control group design**.

A good illustration of this design is an imaginative experiment by Emily Balcetis and David Dunning (2007). These researchers were interested in whether your perception of your physical environment could be altered by your motivation to reduce cognitive dissonance (an uncomfortable psychological state created by cognitive inconsistency). Participants in their first experiment were required to walk between two points on a crowded part of a college campus and estimate the distance walked. After preliminary instructions, participants were handed a bag containing a "Carmen Miranda" costume (consisting of a coconut bra, a grass skirt, and a hat adorned with plastic fruit) to wear while walking between the two points.

The independent variable manipulated was whether participants were given high or low choice to perform the task while wearing the costume. Participants in the high-choice condition were told they could opt for other unspecified tasks, but that the experimenter would prefer that they wear the costume and walk between the two points. Participants in the low-choice condition were told that although other tasks were available, a supervisor had chosen this task for the participants. Participants in the control condition were not given the Carmen Miranda costume or told about alternative tasks. They simply walked between the two points and estimated the distance.

Based on cognitive dissonance theory, Balcetis and Dunning predicted that participants in the low-choice condition would perceive the task as more challenging (because they had to wear the embarrassing costume and had no choice) and consequently perceive the distance walked between the two points as longer than participants in the high-choice and control conditions. As predicted, participants in the low-choice condition estimated longer distances ($M = 182.5$ feet) than participants in the high-choice ($M = 111.1$ feet) condition. Participants in the control group gave distance estimates between these extremes ($M = 161.5$ feet).

QUESTIONS TO PONDER

1. How does a two-group, randomized design work?
2. What are some of the advantages and disadvantages of the two-group, randomized design?
3. How do parametric and nonparametric multigroup, randomized designs work?

Matched-Groups Designs

In some cases, you know or suspect that some subject characteristics correlate significantly with the dependent variable. For example, subjects often differ considerably in their reaction times to simple stimuli. If you were interested in studying the effect of stimulus complexity on reaction time, this large inherent variation in reaction time already present in your subjects could pose a problem. Creating large amounts of error variance could swamp any effect of stimulus complexity, making even large differences in group means statistically unreliable. One way to deal with this problem is to use a matched-groups design.

A **matched-groups design** is one in which matched sets of subjects are distributed at random, one per group, into the groups of the experiment. Figure 10-3 illustrates this process. You begin by obtaining a sample of subjects (group at the top of the figure) from the larger population. Next, you assess the subjects on one or more characteristics that you believe exert an influence on the dependent measure and then group the subjects whose characteristics match. In a reaction-time experiment, for example, participants could be pretested for their simple reaction times and then

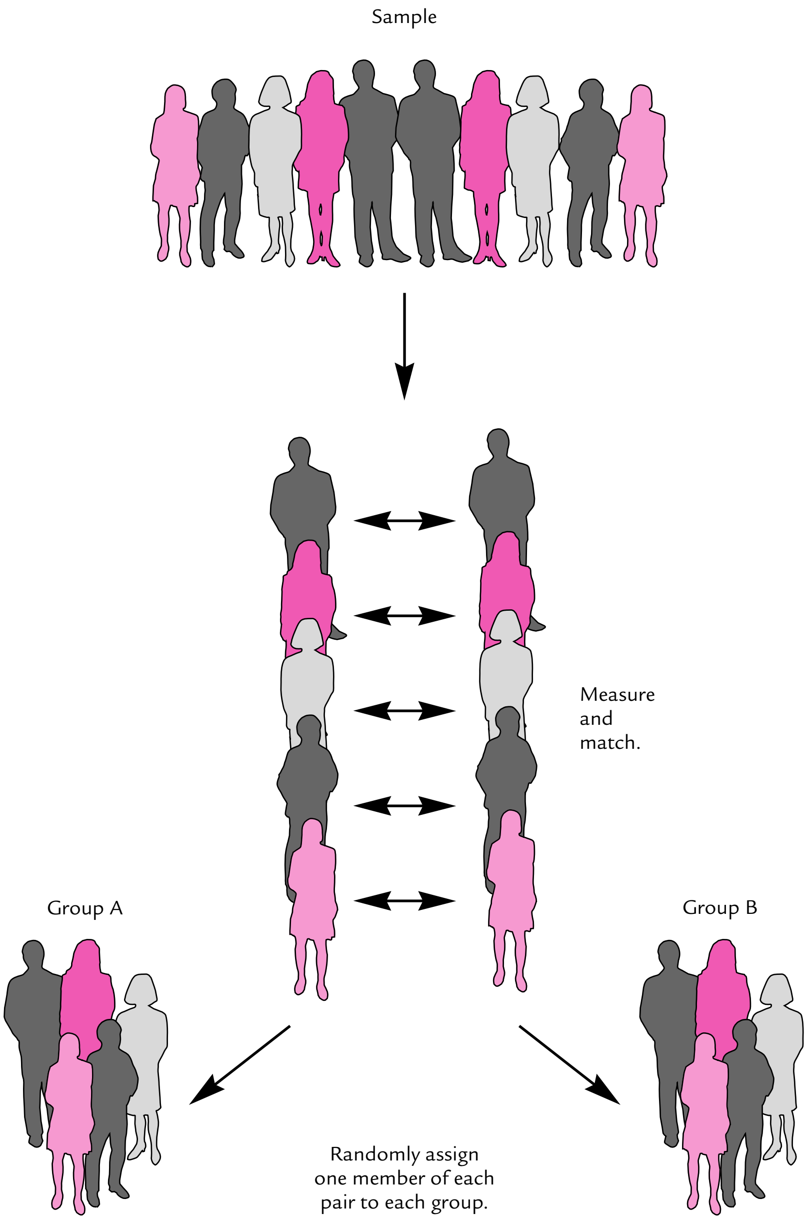


FIGURE 10-3 Matched-groups experimental design with two groups.

grouped into pairs whose reaction times were similar. These pairs of participants are shown in the middle portion of Figure 10-3.

Having matched your participants, you then distribute them randomly across the experimental groups. In the reaction-time experiment, for example, one participant of each pair is randomly assigned to one of the treatments (perhaps to a high-stimulus-complexity condition); the other participant then automatically goes into the other treatment (in this case, a low-stimulus-complexity condition). This assignment to treatments is shown in the bottom of Figure 10-3.

From here on, you conduct the experiment as in the randomized-groups design. You expose your participants to their respective levels of the independent variable and record the resulting data. Then you compare the data from the different groups to determine the effect of the independent variable.

Logic of the Matched-Groups Design Because each of the matched subjects goes into a different group, the effect of the characteristic on which the subjects were matched gets distributed evenly across the treatments. As a result, this characteristic contributes little to the differences between group means. The effect of the error variance contributed by the characteristic has been minimized, making it more likely that any effect of the independent variable will be detected.

Advantages and Disadvantages of the Matched-Groups Design The advantage of matching over random assignment is that it allows you to control subject variables that may otherwise obscure the effect of the independent variable under investigation. Where such variables exist, matching can increase the experiment's sensitivity to the effect of the independent variable (if such an effect is present). This is a potent advantage. You may be able to discover effects that you would otherwise miss. In addition, you may be able to demonstrate a given effect with fewer subjects, thus saving time and money. However, using a matched design is not without risks and disadvantages.

One risk involved in using a matched design concerns what happens if the matched characteristic does *not* have much effect on the dependent variable under the conditions of the study. Matched designs require you to use somewhat modified versions of the inferential statistics you would use in an unmatched, completely randomized design (see Chapter 14). These statistics for matched groups are somewhat less powerful than their unmatched equivalents. This means that they are less able to discriminate any effect of the independent variable from the effect of uncontrolled, extraneous variables.

If the matched characteristic has a *large* effect on the dependent variable, eliminating this effect from group differences will more than compensate for the reduced sensitivity of the statistic, resulting in a more sensitive experiment. However (and this is an important "however"), if the matched characteristic has *little or no effect* on the dependent variable, then matching will do no good. Worse, the loss of statistical power will result in a *reduced* ability to detect the effect of the independent variable. For this reason, use matching only when you have good reason to believe that the matching variable has a relatively strong effect on the dependent measure.

When using a matched design, you also must be sure that the instrument used to determine the match is valid and reliable. If you want to match on IQ, for example, be sure that the test you use to measure IQ is valid and reliable. Of course, for some characteristics, such as race, age, and sex, this is usually not a problem.

In other respects, matched-groups designs have the same advantages and disadvantages as randomized-groups designs. However, the requirement for pretesting and matching makes the matched design more demanding and time consuming than the randomized design. In addition, you may require a larger subject pool if you cannot find a match for certain subjects and must discard them from the study. This may be particularly troublesome if you are attempting to match subjects on more than one variable or if the subject pool is limited.

Any of the randomized-groups designs described in the previous sections of this chapter could be modified into a matched-groups design. The simplest case, described next, involves the two-group design.

The Matched-Pairs Design The **matched-pairs design** is the matched-groups equivalent to the randomized two-group design. The hypothetical reaction time experiment just described uses a matched-pairs design. As with the randomized two-group design, the need for only two groups makes this approach relatively economical of time and subjects but does limit the amount of information you can obtain from the experiment.

Matched-Multigroup Designs The same approach used in the matched-pairs design can be extended to other, more complex designs involving multiple levels of a single factor (single-factor, multigroup designs) or multiple factors (factorial designs). You use these matched-groups designs to gain control over subject-related variables that affect your dependent variable and thus tend to obscure any effects of your independent variable.

Using the matching strategy on these multigroup designs requires you to find a matched subject for every treatment group in your experiment. Thus, if your experiment included four treatment groups, you would need to find quadruplets of subjects having similar characteristics on the variables being matched. After matching subjects, you would distribute the subjects from each quadruplet randomly across your experimental groups.

As you might guess, matching becomes unwieldy if your design has more than about three groups because it becomes increasingly difficult to find three, four, or more subjects with equivalent scores on the variable or variables to be matched. In this case, a better approach might be to use a within-subjects design. The within-subjects design eliminates the need for measuring and matching subject variables, reduces the number of subjects required for the experiment, and yet provides the ultimate degree of matching—in effect, each subject is matched with himself or herself. Unfortunately, situations do occur in which the within-subjects design cannot or should not be used. In such cases, matching may be your best alternative.

QUESTIONS TO PONDER

1. What is a matched-groups design and when would you use one?
2. How does a matched-pairs design differ from a randomized two-group design?
3. What are some of the advantages and disadvantages of the matching strategy?

WITHIN-SUBJECTS DESIGNS

In between-subjects designs, you randomly assign subjects to groups and then expose each group to a different, *single* experimental treatment. You measure each subject's performance on the dependent variable, calculate the average score for each group, and then compare the means to determine whether the independent variable or variables had any apparent influence on the dependent variable. You then subject the data to a statistical analysis to assess the reliability of your conclusions.

The within-subjects design follows the same basic strategy as the between-subjects design with one important difference. In the within-subjects design, each subject is exposed to all levels of your independent variable rather than being randomly assigned to one level. This strategy is shown in Figure 10-4 with a simple two-treatment experiment. Notice that each participant's performance is measured under Treatment A and then again under Treatment B. The design is called *within-subjects* because comparison of the treatment effects involves looking at changes in performance within each participant across treatments. Because participant behavior is measured repeatedly, the within-subjects design is sometimes also called a *repeated-measures design*.

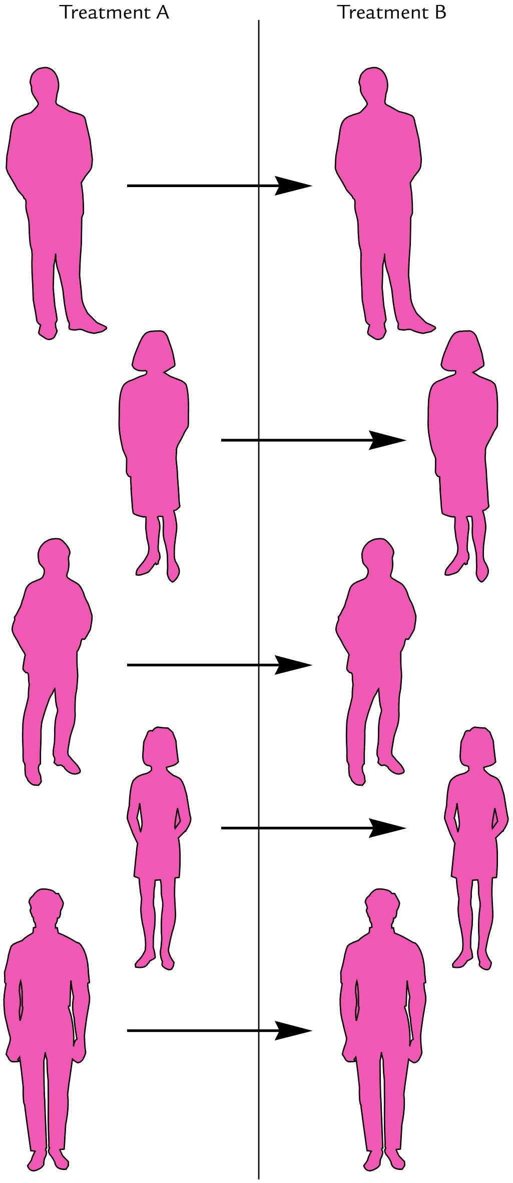
Within-subjects designs are closely related to the matched-groups designs that we discussed in the previous section, in which subjects are first matched into sets on some characteristic (such as IQ score) and then members of each matched set are assigned at random, one to each treatment condition. You might think of a within-subjects design as providing the ultimate in matching because each participant in effect serves as his or her own matched partner across the treatments.

An Example of a Within-Subjects Design: Does Caffeine Keep Us Going?

It is a ritual that is played out in many circumstances at various times. Something has to be done that requires burning the midnight oil (e.g., studying for a final exam or driving all night to your vacation destination). In such circumstances, we often turn to the most widely available stimulant: caffeine. Caffeine is commonly found in coffee and in a variety of caffeinated soft drinks. It is commonly believed that consuming such beverages will keep us awake and keep our cognitive wits sharp while we burn the midnight oil. But is this true? An experiment using a within-subjects design conducted by H. J. Smit and P. J. Rogers (2000) investigated this issue.

Participants in this experiment were 23 adult males and females. Each participant consumed a beverage containing 0, 12.5, 25, 50, or 100 mg of caffeine. Each participant received all five dosages of caffeine. This is what makes this experiment a within-subjects design. Each beverage was administered once per week and the order in which the beverages were consumed was counterbalanced; that is, different participants received the doses in a different order (we discuss this issue in more detail below). Participants completed two measures of cognitive performance (a simple reaction-time measure and a more complex task: identifying when a string of three odd or even digits appeared on a computer screen) before and after exposure to the caffeine. The results revealed a dose-response relationship between caffeine dosage and reaction time. Generally, higher doses of caffeine resulted in faster reaction times. On the more

FIGURE 10-4 Simple two-treatment within-subjects design. Each individual receives both treatments.



complex cognitive task, increasing the dosage of caffeine led to better performance but only among participants who normally consume higher levels of caffeine.

Advantages and Disadvantages of the Within-Subjects Design

Within-subjects designs offer some powerful advantages over the equivalent between-subjects designs if certain conditions can be met. They also introduce problems whose solution adds complexity to the basic designs, and they present other disadvantages as well. We begin by examining the advantages.

Advantages of the Within-Subjects Design Previously in this chapter, we noted that scores *within* a treatment group differ for reasons having nothing to do with your independent variable. These differences arise from the effects of extraneous variables, which include relatively stable subject-related characteristics as well as momentary fluctuations that change each subject's performance from moment to moment. Such error variance can be a serious problem because it may mask any effects of your independent variable. Recall that a major strategy for dealing with error variance in the between-subjects design is to randomly assign subjects to treatment groups and to apply statistical analysis to your data to estimate the probability with which chance alone could have produced the effect. When subject-related factors are large, they exert a strong influence on performance, resulting in levels of error variance that obscure the effect of your independent variable. Matching can help reduce this important source of error variance.

The within-subjects design pushes the logic of matching to the limit. Each subject is matched with other subjects who are virtual clones of each other, because they are in fact the same subject. All subject-related factors (such as age, weight, IQ, personality, religion, and gender) are literally identical across treatments. Thus, any performance differences across treatments cannot be due to error variance arising from such differences, as is the case in the between-subjects design. Because of the reduced error variance, the within-subjects design is more powerful (i.e., more sensitive to the effects of your independent variable) than the equivalent between-subjects design. Thus, you are more likely to detect an effect of your independent variable. A second benefit of this increased power is that you can use fewer subjects in your experiment. For example, a four-group between-subjects design with 10 subjects per treatment would require 40 subjects. The equivalent within-subjects experiment would require only 10 subjects, representing a significant savings in time, materials, and money. For example, in the Smit and Rogers (2000) study on caffeine, only 23 participants were required. Of course, you could always use more subjects in a within-subjects design if you needed extra power for your statistical analysis to detect the effect of a weak independent variable.

Disadvantages of the Within-Subjects Design Although the within-subjects design has its advantages, it also has some important disadvantages, which may preclude its use in certain situations. One disadvantage is that a within-subjects design is more demanding on subjects because each subject must be exposed to every level of the experimental treatment. A complex design involving, for example, nine treatments would require a great deal of time to complete. It may be difficult to find participants willing to take part in such an experiment. Those who do take part may become bored or fatigued after being in an experiment that might be several hours long. You can get around the problem of fatigue and boredom by administering only one or two treatments per session, spreading sessions out over some period of time. However, if you take this approach, you may lose some participants from the experiment because they fail to show up for one or more sessions.

Subject attrition also can occur if you make a mistake while administering one of your treatments (e.g., you read the wrong instructions), if you experience equipment failure, or (in the case of animal research) if your subject dies. In each case, you have to throw out the data from the lost subjects and start over.

A second and potentially more serious problem with the within-subjects design is its ability to produce carryover effects. **Carryover effects** occur when a previous treatment alters the behavior observed in a subsequent treatment. The previous treatment changes the subject, and those changes carry over into the subsequent treatment, in which they change how the subject performs. This upsets the "perfect match" of subject characteristics that the within-subjects design is supposed to provide.

As an illustration of carryover effects, imagine that you are conducting an experiment to assess the effect of two kinds of practice (simple rehearsal and rehearsal plus imagery) on memory for lists of concrete nouns. Your participants first learn a list of nouns using simple rehearsal and then are tested for retention. Next the participants learn a second list of nouns, using rehearsal plus imagery, and are again tested. You find that participants correctly recall more nouns when they used a rehearsal-plus-imagery technique than when they used rehearsal alone. However, you cannot confidently conclude that the former technique is superior to the latter.

The problem is that the rehearsal-alone treatment gave participants practice memorizing nouns. They may have done better in the rehearsal-plus-imagery treatment simply because they were more practiced at the task rather than because of any effect of imagery. The previous exposure to the rehearsal-alone treatment may have changed the way participants performed in the subsequent treatment.

Carryover effects can be a serious problem in any within-subjects design. Between-subjects designs do not suffer from carryover effects simply because there are no previous conditions from which effects can carry over. A matched-groups design may provide a reasonable compromise in those situations in which carryover is a serious problem but in which you want to retain the control over subject variables provided by a within-subjects design.

The problem of carryover in within-subjects designs has received plenty of attention from researchers, who have developed strategies to deal with it. The next section identifies potential sources of carryover. After that, we describe several design options that can help you deal with potential carryover effects.

QUESTIONS TO PONDER

1. How does a within-subjects design differ from a between-subjects design?
2. What are the advantages of the within-subjects experimental design?
3. What are the disadvantages of the within-subjects experimental design?
4. How do carryover effects influence the interpretation of the results from a within-subjects experiment?

Sources of Carryover

Carryover effects can arise from a number of sources, including the following:

- *Learning*. If a subject learns how to perform a task in the first treatment, performance is likely to be better if the same or similar tasks are used in subsequent treatments. For example, rats given alternate sessions of reinforcement and extinction show faster acquisition of lever pressing across successive reinforcement sessions and more rapid return to baseline rates of responding across successive extinction sessions.

- *Fatigue*. If performance in earlier treatments leads to fatigue, then performance in later treatments may deteriorate, regardless of any effect of the independent variable. If measuring your dependent variable involves having participants squeeze against a strong spring device to determine their strength of grip, for example, the participants are likely to tire if repeated testing takes place over a short period of time.
- *Habituation*. Under some conditions, repeated exposure to a stimulus leads to reduced responsiveness to that stimulus because the stimulus is becoming more familiar or expected. This reduction is termed *habituation*. Your subjects may jump the first time you surprise them with a sudden loud noise, but they may not do so after repeated presentations of the noise.
- *Sensitization*. Sometimes exposure to one stimulus can cause subjects to respond more strongly to another stimulus. In a phenomenon called *potentiated startle*, for example, a rat will show an exaggerated startle response to a sudden noise if the rat has recently received a brief foot shock in the same situation.
- *Contrast*. Because of contrast, exposure to one condition may alter the responses of subjects in other conditions. If you pay your participants a relatively large amount for successful performance on one task and then pay them less (or make them work harder for the same amount) in a subsequent task, they may feel underpaid. Consequently, they may work less than they otherwise might have. This change occurs because subjects can compare (contrast) the treatments.
- *Adaptation*. If subjects go through a period of adaptation (e.g., becoming adjusted to the dark), then earlier results may differ from later results because of the adaptive changes. Adaptive changes may increase responsiveness to a stimulus (e.g., sight gradually improves while you sit in a darkened theater) or decrease responsiveness (e.g., you readjust to the light as you leave the theater). Adaptation to a drug schedule is a common example. If adaptation to the drug causes a reduced response, the change is called *tolerance*.

Dealing With Carryover Effects

You can deal with carryover effects in three ways: You can (1) use counterbalancing to even out carryover effects across treatments, (2) take steps to minimize carryover, and (3) separate carryover effects from treatment effects by making treatment order an independent variable.

Counterbalancing In **counterbalancing** you assign the various treatments of the experiment in a different order for different subjects. The goal is to distribute any carryover equally across treatments so that it does not produce differences in treatment means that could be mistaken for an effect of the independent variable. Smit and Rogers (2000) used this strategy in the experiment presented earlier. Recall that each participant received the various caffeine doses in a different, counterbalanced order.

Two counterbalancing options are complete counterbalancing and partial counterbalancing. *Complete counterbalancing* provides every possible ordering of treatments and assigns at least one subject to each ordering. Table 10-2 shows an example of a completely counterbalanced single-factor design that includes three treatments. Six subjects are to be tested (identified as subjects S_1 through S_6), one for each possible ordering of treatments T_1, T_2, and T_3. Note that in a completely counterbalanced design, every treatment follows every other treatment equally often across subjects, and every treatment appears equally often in each position (first, second, etc.).

The minimum number of subjects required for complete counterbalancing is equal to the number of different orderings of the treatments: k treatments have exactly $k!$ (k factorial) orders, where $k! = k\,(k - 1)(k - 2) \cdots (1)$. For example, with three treatments (as in our example), the number of treatment orders is $3 \times 2 \times 1 = 6$. If you need to increase the number of subjects in order to improve statistical power, add the same number of additional subjects to each order so that the number of subjects receiving each order remains equal.

Complete counterbalancing is practical for experiments with a small number of treatments, but this approach becomes increasingly burdensome as the number of treatments grows. For an experiment using only four treatments, the $4 \times 3 \times 2 \times 1 = 24$ possible treatment orders require at least 24 subjects to complete the counterbalancing. The economy of subjects that makes a within-subjects approach attractive erodes rapidly.

Fortunately, you can recover some of this economy by switching to the second type of counterbalancing. *Partial counterbalancing* includes only some of the possible

TABLE 10-2 Counterbalanced Single-Factor Design With Three Treatments

	TREATMENTS		
Subjects	T_1	T_2	T_3
S_1	1	2	3
S_2	1	3	2
S_3	2	1	3
S_4	2	3	1
S_5	3	1	2
S_6	3	2	1

treatment orders. The orders to be retained are chosen randomly from the total set with the restriction that each treatment appear equally often in each position. Table 10-3 displays all 24 possible orders for a four-treatment experiment, followed by a subset of 8 randomly selected orders that meet this criterion.

When you use partial counterbalancing, you assume that randomly chosen orders will randomly distribute carryover effects among the treatments. Although carryover effects may not balance out under such conditions, they usually will come close to doing so. Furthermore, the likelihood that treatments will differ because of carryover can be evaluated statistically and held to an acceptable level. If you choose to make the number of treatment orders in your partially counterbalanced design equal to the

TABLE 10-3 Twenty-Four Possible Treatment Orders for a Four-Treatment Within-Subjects Design and a Randomly Selected Subset in Which Each Treatment Appears Equally Often in Each Position

ENTIRE SET OF TREATMENT ORDERS	SELECTED SUBSET
1. ABCD	1. DABC
2. ABDC	2. ABCD
3. ACBD	3. CDAB
4. ACDB	4. BCDA
5. ADBC	5. DCBA
6. ADCB	6. ADCB
7. BACD	7. BADC
8. BADC	8. CBAD
9. BCAD	
10. BCDA	
11. BDAC	
12. BDCA	
13. CABD	
14. CADB	
15. CBAD	
16. CBDA	
17. CDAB	
18. CDBA	
19. DABC	
20. DACB	
21. DBAC	
22. DBCA	
23. DCAB	
24. DCBA	

number of treatments, you can use a *Latin square design* to ensure that each treatment appears an equal number of times in each ordinal position. For more information on how to construct a Latin square design, see Edwards (1985).

Counterbalancing (whether complete or partial) can be counted on to control carryover only if the carryover effects induced by different orders are of the same approximate magnitude. Consider the case of a simple two-treatment experiment shown in Table 10-4. This case has only two possible orders: 1→2 and 2→1. Assume that carryover from Treatment 1 to Treatment 2 increases the mean score of Treatment 2 by 10 points and that carryover from Treatment 2 to Treatment 1 has a similar effect on the mean score of Treatment 1. Table 10-4 shows the result for a completely counterbalanced design. Note that the two carryover effects, being equal, cancel out each other. When carryover effects are equivalent across orders, counterbalancing is effective.

In contrast, when the magnitude of the carryover effect differs for different orders of treatment presentation, counterbalancing may be ineffective. Table 10-5 illustrates this problem known as *differential carryover effects* (Keppel, 1982). In this example, the carryover from Treatment 2 to Treatment 1 averages 20 points—twice the carryover from Treatment 1 to Treatment 2. Thus, you have a treatment-by-position interaction. When this occurs, no amount of counterbalancing will eliminate the carryover effects (Keppel, 1982).

The most serious asymmetry in carryover effects occurs when a treatment produces *irreversible changes*. The classic type of irreversible change is that produced by a treatment such as brain lesioning. The effects of the operation, once present, cannot be undone. A somewhat less serious change may occur if subjects learn to perform a task in one treatment, and this learning then alters the way in which they perform in a subsequent treatment. It may not be possible to restore subjects to the "naive" state once they have learned the task. In either case, you would want to choose a between-subjects approach.

Taking Steps to Minimize Carryover The second way to deal with carryover effects is to try to minimize or eliminate them. Of course, you would want to do this only if the carryover effects were not themselves the object of study. Minimizing carryover effects reduces error variance and increases the power of the design.

TABLE 10-4 Balancing of Order Effects in a Counterbalanced Two-Treatment Design

	TREATMENT		
	1	***2***	
Actual treatment effect	40	30	(difference = 10)
Carryover effect (1→2)		10	
Carryover effect (2→1)	10	—	
Observed treatment effect	50	40	(difference = 10)

TABLE 10-5 Failure of Order Effects to Balance Out in a Counterbalanced Two-Treatment Design

	TREATMENT		
	1	*2*	
Actual treatment effect	40	30	(difference = 10)
Carryover effect (1–2)		10	
Carryover effect (2–1)	20	—	
Observed treatment effect	60	40	(difference = 20)

Not all sources of carryover can be minimized. For example, permanent changes produced by learning inevitably carry over into subsequent treatments and affect behavior. You cannot return your subjects to the naive state in preparation for a second treatment. However, if you are not interested in the effect of learning per se, you may be able to pretrain your subjects before introducing your experimental treatments. Psychophysical experiments (testing such things as sensory thresholds) and experiments on human decision making often make use of such "practice sessions" to familiarize participants with the experimental tasks. The practice brings their performances up to desired levels, where they stabilize, and effectively eliminates changes caused by practice as a source of carryover.

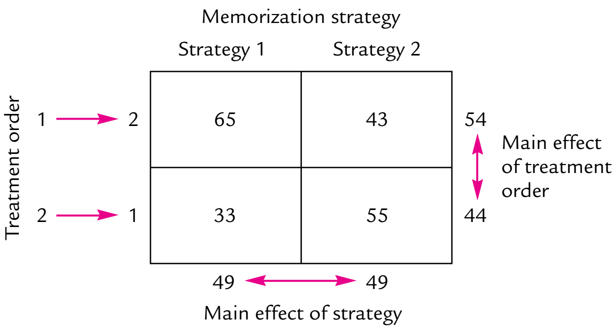
Adaptation and habituation changes can be dealt with similarly. Before introducing the treatments, allowing time for subjects to adapt or habituate to the experimental conditions can eliminate carryover from these sources.

Another way to deal with habituation (if habituation is short term), adaptation, and fatigue is to allow breaks between the treatments. If sufficiently long, the breaks allow subjects to recover from any habituation, adaptation, or fatigue induced by the previous treatment.

You can take steps to minimize carryover effects in combination with either of the other two strategies. If you simply want to control carryover, you could take these steps and then use counterbalancing to distribute whatever carryover remains across treatments. Similarly, if you want to determine whether certain variables contribute to carryover, you could take steps to minimize other potential sources of carryover and then treat the variables of interest as independent variables, as described in the following section.

Making Treatment Order an Independent Variable A third way to deal with the problem of carryover is to make treatment order an independent variable. Your experimental design will expose different groups of subjects to different orderings of the treatments, just as in ordinary counterbalancing. However, you include a sufficient number of subjects in each group to permit statistical analysis of treatment order as

FIGURE 10-5 A design in which order of treatments is made an independent variable.



a separate independent variable. For example, if you were going to conduct a one-factor experiment to compare the effect of two memorization strategies on recall, you could design the experiment to include the order of testing as a second independent variable. Figure 10-5 illustrates the resulting design, which now includes *two* independent variables. Called a *factorial design*, it requires a special type of analysis to separately evaluate the effect of each and is discussed later in this chapter (see "Factorial Designs: Designs with Two or More Independent Variables"). For now, it is enough to know that this design allows you to separate any carryover effects from the effect of your experimental treatment.

The main advantage of making order of treatments an independent variable is that you can measure the size of any carryover effects that may be present. You can then take these effects into account in future experiments. If you find that carryover is about equal in magnitude regardless of the order of treatments, for example, then you can be confident that counterbalancing will eliminate any carryover-induced bias.

In addition to identifying carryover effects, the strategy of making treatment order an independent variable provides a direct comparison of results obtained in the within-subjects design with those obtained in the logically equivalent between-subjects design. This comparison can be made because every treatment occurs first in at least one of the treatment orders. These "first exposures" provide the data for a purely between-subjects comparison in the absence of carryover effects.

Grice (1966) notes that between-subjects and within-subjects designs applied to the same variables do not always produce the same functional relationships. The reason is that subjects in the within-subjects experiment are able to make comparisons across treatments whereas those in the supposedly equivalent between-subjects experiment are not. Imagine, for example, a study in which participants must rate the attractiveness of pictures on a 5-point scale. In one version of the study, different groups of participants see only one of the pictures. In another version, each participant views all the pictures. A particular picture may rate, say, 5 on the scale when viewed by itself. However, when seen in the context of the other pictures, the same picture may look better (or worse) in comparison and thus may produce a different rating. Such changes in response that arise from comparison, termed *contrast effects*, are possible only in the within-subjects version of the study.

The presence of such effects in some designs, but not in others, often can explain why studies manipulating the same variables sometimes yield different results.

Although making treatment order a factor in your experiment can provide important information about the size of carryover effects and can pinpoint the source of differences between findings obtained from within-subjects versus between-subjects experiments, the technique does have disadvantages. Every treatment order requires a separate group of subjects. These subjects must be tested under every treatment condition. The result is a complex, demanding experiment that is costly in terms of numbers of subjects and time to test them. Furthermore, these demands escalate rapidly as the number of treatments (and therefore number of treatment orders) increases. This latter problem is the same one encountered when using completely counterbalanced designs. For these reasons, the approach is practical only with a small number of treatments.

QUESTIONS TO PONDER

1. What are the sources of carryover effects in a within-subjects design?
2. Under what conditions will counterbalancing be effective or ineffective in dealing with carryover effects?
3. When do you use a Latin square design?
4. What strategies can be used to deal with carryover effects?

When to Use a Within-Subjects Design

Given the problems created by the potential for carryover effects, the best strategy may be to altogether avoid within-subjects designs. If you decide to do this, you have a lot of company. However, you should not let these difficulties prevent you from adopting the within-subjects design when it is clearly the best approach. There are several situations in which the within-subjects design is best, and others in which it is the *only* approach.

Subject Variables Correlated With the Dependent Variable You should strongly consider using a within-subjects design when subject differences contribute heavily to variation in the dependent variable. As an example, assume that you want to assess the effect of display complexity on target detection in a simulated air traffic–controller task. Your display simulates the computerized radar screen display seen in actual air traffic–control situations with aircraft (the targets) appearing as blips in motion across the screen. Your independent variable is the amount of information being displayed (the dots alone or dots plus transponder codes, altitude readings, etc.). Because this is a pilot study (no pun intended), you are using college sophomores as participants rather than real air traffic controllers.

Your student participants are likely to differ widely in their native ability to detect targets, regardless of the display complexity. If you were to conduct the experiment using a between-subjects design, this large variation in ability would contribute greatly to within-group error variance. As a consequence, the between-group differences would probably be obscured by these uncontrolled variations.

In this case, you could effectively eliminate the impact of subject differences by adopting a within-subjects design. Each participant is exposed to every level of display complexity. Because each participant's native ability at target detection remains essentially the same across all the treatment levels, the changes (if any) in target detection across treatments would clearly stand out. Of course, you would have to be reasonably sure that practice at the task does not contribute to success at target detection (or at least that the effect of such practice could be distributed evenly across treatments by using counterbalancing) before you decided to adopt the within-subjects approach. One way to eliminate practice as a source of confounding would be to include several practice sessions in which your participants became proficient enough at the task that little further improvement would be expected.

Economizing on Subjects You also should consider using a within-subjects design when the number of available subjects is limited and carryover effects are absent (or can be minimized). If you were to use actual air traffic controllers in the previous study, for example, you probably would not have a large available group from which to sample. You probably would not be able to obtain enough participants for your study to achieve statistically reliable results using a between-subjects design. Using a within-subjects design would reduce the number of participants required for the study while preserving an acceptable degree of reliability.

Assessing the Effects of Increasing Exposure on Behavior In cases in which you wish to assess changes in performance as a function of increasing exposure to the treatment conditions (measured as number of trials, passage of time, etc.), the within-subjects design is the only option (unless you have enough control to use a single-subject design—see Chapter 12). Designs that repeatedly sample the dependent variable across time or trials are frequently used in psychological research to examine the course of processes such as reinforcement, extinction, fatigue, and habituation. These changes occur as a function of earlier exposure to the experimental conditions and thus represent carryover effects. However, the carryover effects in these designs are the object of study rather than something to be eliminated by measures such as counterbalancing.

Be aware, however, that not all carryover effects can or should be studied within the framework of a within-subjects design. For example, transfer-of-training studies (in which the effects of previous training on later performance of another task are assessed) are not good candidates for a within-subjects approach. This is because the earlier training may have effects on later performance that cannot easily be reversed. For example, if you wanted to compare performance on a mirror-tracing task with and without previous training, subjects first receiving the "previous training" condition probably could not be brought back to the "naive" state prior to being given the "no previous training" condition. In this case, you would have to use separate groups for the training and no-training conditions.

Within-Subjects Versus Matched-Groups Designs

The within-subjects and matched-groups designs both deal with error variance by attempting to control subject-related factors. As we have seen, the two designs go about this in different ways. In the matched-groups design you measure subject variables and match subjects accordingly, whereas in the within-subjects design you use the same subjects in all treatments. Both designs take advantage of any correlations between subject variables and your dependent variable to improve power, and both use similar statistical analyses to take this correlation into account. However, if the correlation between those subject variables and the dependent variable is weak, a randomized-groups design will be more powerful. Thus, if you have reason to believe that the relationship between subject variables and your dependent variable is weak, use a randomized-groups design.

A matched-groups design would be a better choice than a within-subjects design if you are concerned that carryover effects will be a serious problem. Although you lose the economy of having fewer subjects, you avoid the possibility of carryover effects while preserving the power advantage made possible by matching.

Types of Within-Subjects Designs

Just as with the between-subjects design, the within-subjects design is really a family of designs that incorporate the same basic structure. This section discusses several variations on the within-subjects design. These variations include the single-factor, multilevel within-subjects design (in both parametric and nonparametric versions); the multifactor within-subjects design; and multivariate within-subjects designs.

The Single-Factor Two-Level Design The *single-factor two-level design* is the simplest form of within-subjects design and includes just two levels of a single independent variable. All subjects receive both levels of the variable, but half the subjects receive the treatments in one order and half in the opposite order. The scores within each treatment are then averaged (ignoring the order in which the treatments were given), and the two treatment means are compared.

This design is directly comparable to the two-group between-subjects design while offering the general advantages and disadvantages of the within-subjects approach. If order effects are not severe and are approximately equal for both orders, then counterbalancing will control the order effects without introducing excessive error variance. If the dependent variable is strongly affected by subject-related variables, then the two-factor within-subjects design will control this source of variance, and the experiment will more likely detect the effect (if any) of the independent variable. However, if the dependent variable is not strongly affected by subject-related variables, this design will be less effective in detecting the effect of the independent variable than will its two-group between-subjects equivalent.

Single-Factor Multilevel Designs Just as with the between-subjects design, the within-subjects design can include more than two levels of the independent variable. In the *single-factor multilevel within-subjects design*, a single group of subjects is exposed to three or more levels of a single independent variable. If the independent variable is not a cumulative factor (such as practice), then the order of treatments is counterbalanced to prevent any carryover effects from confounding the effects of the treatments.

TABLE 10-6 Structure of a Counterbalanced Single-Factor Within-Subjects Design With Four Treatments

	TREATMENTS			
Subjects	T_1	T_2	T_3	T_4
S_1	4	1	2	3
S_2	1	2	3	4
S_3	3	4	1	2
S_4	2	3	4	1
S_5	4	3	2	1
S_6	1	4	3	2
S_7	2	1	4	3
S_8	3	2	1	4

Table 10-6 shows the organization of a single-factor within-subjects design with four levels or treatments and eight subjects. In this example, subjects have been randomly assigned to eight different treatment orders with the restriction that each treatment appears equally often in each ordinal position. Each row indicates the ordinal position of each treatment for a given subject. Treatment orders were determined by constructing two Latin squares, one for each batch of four subjects.

Earlier, we distinguished among several types of single-factor between-subjects design, including parametric, nonparametric, and multiple control group versions. The same distinctions can be applied to within-subjects designs. Because the basic features of these designs have been described before, a separate section is not provided here for each type. Instead, a parametric within-subjects experiment will illustrate the single-factor multilevel within-subjects design. Peterson and Peterson (1959) conducted a now-classic study of memory processes. This study was designed to determine the effect of retention interval on memory for three-consonant trigrams (such as *JHK*). Retention intervals of 3, 6, 9, 12, 15, and 18 seconds were tested, with each participant receiving all the retention intervals and with order of intervals counterbalanced across participants.

To prevent them from rehearsing the trigram during the retention intervals, the participants were kept busy doing a demanding mental arithmetic task. Figure 10-6 shows the results. Probability of correct recall was found to decline sharply as the retention interval increased. This evidence provided strong support for the existence of short-term memory as a separate entity from long-term memory.

The Petersons' (1959) experiment could have been conducted using a between-subjects design. In this case, however, the use of the within-subjects design reduced the time and the number of participants required to complete the study. At the same time, this design prevented the sometimes large individual differences in recall performance from obscuring the effect of retention interval.

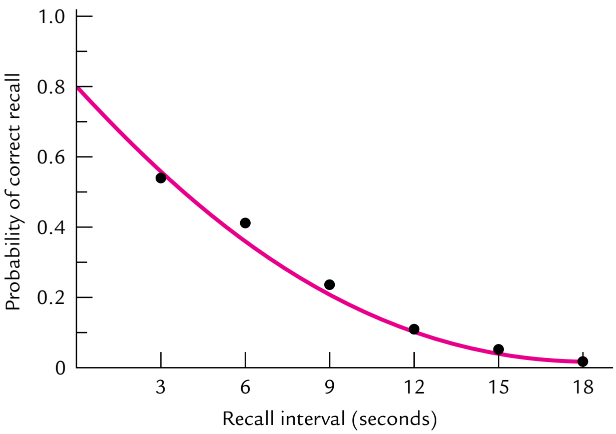


FIGURE 10-6 Probability of correct recall as a function of retention interval.

SOURCE: Peterson and Peterson, 1959; reprinted with permission.

The results from a multilevel within-subjects design are interpreted much like the results from a multigroup between-subjects design. If the independent variable is quantitative (i.e., measured along an interval or ratio scale), as in the Peterson and Peterson (1959) study or as in designs in which trials or time is the independent variable, then the design is said to be *parametric* (just as with between-subjects designs). In that case, your primary interest in conducting the study may be to determine the form of the function relating the independent and dependent variables. However, if your independent variable represented different categories (such as different types of drugs), then talking about functional relationships would be meaningless. In that case, you would want to compare the effects of the different treatments with each other and with those of any control conditions included in the design.

QUESTIONS TO PONDER

1. Do between-subjects and within-subjects designs applied to the same variables always produce the same functional relationship? Why or why not?
2. When should you consider using a within-subjects design instead of a between-subjects design?
3. When should you consider using a matched-groups design rather than a within-subjects design?
4. How do single-factor and multifactor experimental design differ?

FACTORIAL DESIGNS: DESIGNS WITH TWO OR MORE INDEPENDENT VARIABLES

Thus far, the discussion has considered designs involving only one independent variable being manipulated in an experiment. If you wanted to assess the effects of several independent variables on a given dependent variable, one solution would be to conduct a separate experiment for each independent variable of interest. Although this approach is sometimes used, you may be able to gain more information at less expense by using a **factorial design** that incorporates two or more independent variables in a single experiment. Let's take a look at an example of a factorial design before we discuss this design in detail.

An Example of a Factorial Design: Can That Witness Really *Not* Remember an Important Event?

In January 2007, Lewis "Scooter" Libby, former Vice President Cheney's chief of staff, was convicted of perjury and obstruction of justice relating to the leaking of the name of a CIA operative to the press. Libby claimed that he could not remember revealing the operative's name, a claim that the jury rejected. Apparently, members of the jury could not understand how someone could fail to remember such an important event. Is there any credibility to Libby's statement that he could not recall having leaked the name? Apparently, there is.

The problem that Libby faced was that revealing the name took place at a time when the event had not yet assumed importance. Only after the story about the CIA operative's name was sensationalized in the press did the event take on importance.

Could it be that Libby (and countless other witnesses in other cases) really didn't recall revealing the name because he attached no importance to the event at the time? Karim Kassam, Daniel Gilbert, Jillian Swencionis, and Timothy Wilson (2009) addressed this question in an experiment that used a between-subjects factorial design. Participants were shown a number of photographs supposedly taken from a high school yearbook. Each photograph was accompanied by five facts about the person pictured. Participants were randomly assigned to the role of "memorizer" or "judge." Memorizers were told to study the five facts before they saw the photographs. Further, they were told that they would receive 10¢ for each fact correctly recalled. Before they studied the materials, memorizers were randomly assigned to one of three memorization conditions varying the motivation to recall (MTR). Participants in the MTR-at-encoding condition were told *before* they studied the material that they would receive a 50¢ bonus for each fact they could recall about Beryl White (one of the people shown in one of the photographs). Participants in the MTR-at-retrieval condition were told about the 50¢ bonus *after* they studied the material. Participants in the control group (no-MTR condition) were not told of any bonus. Participants assigned to the "judge" condition did not study the material. Instead, judges were read the instructions from the MTR-at-encoding, MTR-at-retrieval, or no-MTR conditions and told to predict the number of facts that participants would recall under each condition.

Before we look at what Kassam et al. found, let's analyze the experimental design they used. Because different participants were randomly assigned to each condition in the experiment, the design is between-subjects. The inclusion of the two independent variables (role of the participant and memorization condition) makes the design a two-factor, between-subjects design. Finally, because there were three levels of how material was learned by memorizers and two levels of the role to which participants were assigned, the design is a 3 (MTR-at-encoding, MTR-at-retrieval, or no-MTR) × 2 (memorizer or judge) between-subjects design. Figure 10-7 shows the Kassam et al. experimental design graphically.

The results showed effects of both memorization and role condition. There was also a significant interaction (shown in Figure 10-8), which showed that memorizers recalled more facts about Beryl White (M = 3.83) in the MTR-at-encoding condition than in either the MTR-at-retrieval (M = 2.38) or no-MTR condition (M = 1.91). The number of facts predicted by judges who received the MTR-at-encoding instructions (M = 3.72) did not differ significantly from those who received the MTR-at-retrieval instructions (M = 3.49). Judges who received the no-MTR instructions predicted that fewer facts would be recalled (M = 2.43) than the previous two conditions.

So, what do these results tell us about Scooter Libby's dilemma? It is clear that Libby's contention that he could not recall mentioning the operative's name before the story broke has some validity. Look at the Kassam et al. results again. Participants

		Role	
		Memorizer	Judge
MTR Condition	MTR at encoding		
	MTR at retrieval		
	No MTR		

FIGURE 10-7 A 3 × 2 factorial design investigating participant role and motivation to recall (MTR).
SOURCE: Based on information in Kassam, Gilbert, Swencionis, & Wilson (2009).

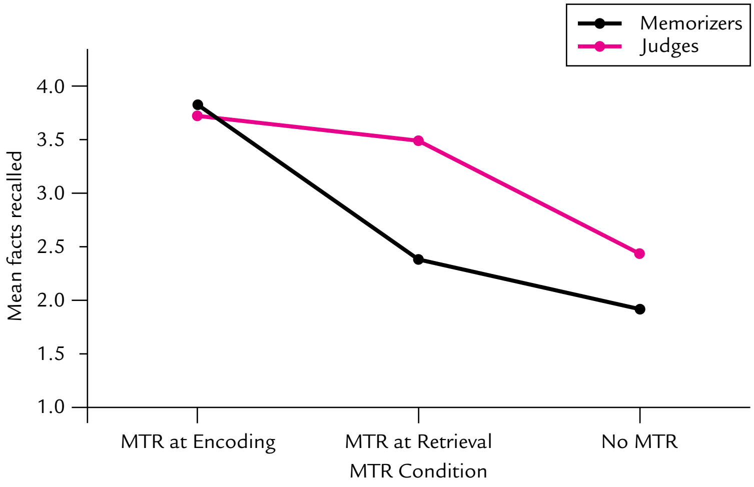


FIGURE 10-8 Interaction between MTR condition and participant role.
SOURCE: Based on data from Kassam, Gilbert, Swencionis, & Wilson (2009).

who were told that one of the pictures (and accompanying facts) were important (the 50¢ bonus) at the time they learned the facts had better recall than those in the condition where importance was attached to the facts only *afte*r learning (the MTR-at-retrieval condition). It is also evident that the jury's skepticism about Libby's memory also has empirical support. It did not matter whether judges received the MTR-at-encoding or MTR-at-retrieval instructions. In both cases, judges expected memory to be good. The lesson from this experiment is clear: There is a disconnect between how memory actually works and how people think it *should* work.

Main Effects and Interactions

On the surface, the Kassam et al. (2009) experiment appears horribly confounded because two independent variables have been allowed to vary at once. This is not the case, however. Because all possible combinations of the levels of the independent variables are represented, you can statistically separate the effects of the independent variables. In fact, the main advantage of a factorial design is that it allows you to assess the effect of each independent variable on the dependent variable separately. So, you can assess whether MTR condition or participant role separately affects recall. These separate effects are known as *main effects*. You can also assess whether a complex relationship exists between your independent variables. That is, you can assess whether MTR condition affects recall for one role (e.g., memorizer) but not the other (e.g., judge). This complex relationship is called an *interaction*.

Main Effects As we just noted, the separate effect of each independent variable is termed a **main effect**. Here's how you calculate the main effects of your independent variables. First, average the group means in the first column and write the result under the first column. Then do the same for the group means in the second column. These two numbers are your *column means*. Now average the group means across the first row and write the result to the right of the first row. Then do the same for the second row. These two numbers are your *row means*. The result should look like Figure 10-9.

Compare the column means. These represent the main effect of participant role, averaged over the three MTR conditions. They are directly analogous to the two means that you would get in a simple two-group design employing these two levels of participant role. Now compare the row means. These represent the main effect of MTR condition, averaged over the two levels of participant role. They are directly analogous to the means of a simple three-group experiment varying MTR condition alone.

A reliable difference in the column means would indicate an effect of participant role, independent of MTR condition. Similarly, a reliable difference in the row means would indicate an effect of MTR condition, independent of role.

Interactions Although the main effects of the independent variables are of considerable interest, they are not the only information that you can extract from a factorial experiment. Nor are they the most interesting. You also can test for the presence of an

MTR Condition	Role		
	Memorizer	Judge	
MTR at encoding	3.83	3.72	3.76
MTR at retrieval	2.38	3.49	2.94
No MTR	1.91	2.43	2.17
	2.88	3.04	

FIGURE 10-9 Organization of a 3 × 2 factorial design with cell means and row and column means.

SOURCE: Based on data from Kassam, Gilbert, Swencionis, & Wilson (2009).

interaction among your independent variables. An **interaction** is present when the effect of one independent variable changes across the levels of another independent variable. For example, Kassam et al. (2009) found that the memorization instructions affected judges and memorizers differently. Memorizers in the MTR-at-encoding condition recalled more facts about the critical target person than memorizers in the MTR-at-retrieval condition. However, judges did not distinguish between these two memorization instructions when predicting how well a memorizer could recall facts.

Figure 10-8 shows the interaction graphically. The differentially drawn lines (colored or black) indicate the groups for which participant role was the same: the colored line for judges and the black line for memorizers. Each line connecting the symbols in Figure 10-8 indicates the effect of memorization instruction at each level of role. These lines are said to represent the **simple main effects** of memorization instruction on the number of facts recalled (or predicted to be recalled). In a two-factor design, a simple main effect represents the effect of one independent variable (e.g., memorization instruction) at a given level of the other independent variable (e.g., role assigned). In general, if the lines of the graph representing different levels of an independent variable are *not parallel*, an interaction *may* be present.

The "may" in the preceding statement results from the fact that the lines drawn on the graph may appear to be nonparallel because of random variability in the data. To determine that an interaction exists, you must establish that the apparent nonparallelism of the lines is not likely to have resulted simply from sampling error. Fortunately, statistical tests are available that simultaneously determine the probable reliability of both the main effects and interactions of a particular experiment. (These tests are discussed in Chapter 14.) Incidentally, if the lines on the graph *are* parallel, no interaction exists, and no statistical test will find one.

Figure 10-10 illustrates several ways that a 2 × 2 factorial experiment might come out. In panel (a), only Factor A has a systematic effect on the dependent variable. In panel (b), only Factor B has an effect. Panel (c) shows an interaction between Factors A and B but no main effect of either factor due to the fact that the average value of each factor, collapsed over the other, is the same for each level. Both Factors A and B affect the dependent variable, but their effects in this case show up only in the interaction. Panel (d) shows a main effect of both factors but no interaction (note the parallel lines).

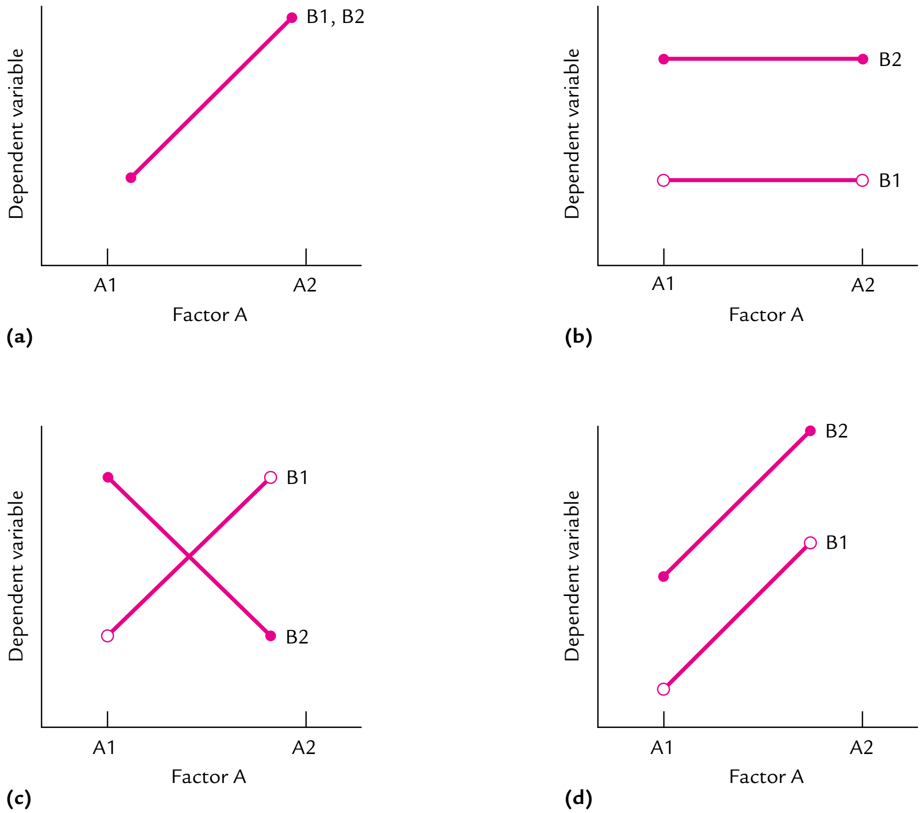


FIGURE 10-10 Some possible outcomes of a 2 × 2 factorial experiment (see text for description).

Factorial Within-Subjects Designs

In the factorial within-subjects design, each subject is exposed to every combination of levels of all the factors (independent variables).

An experiment conducted by Marc Berman, John Jonides, and Stephen Kaplan (2008) illustrates the factorial within-subjects design. Berman et al. wondered whether interacting with nature had greater cognitive benefit than interacting with an urban environment. To evaluate the hypothesis that interacting with nature would have more cognitive benefits than interacting with an urban environment, Berman et al. conducted a 2 × 2 within-subjects experiment.

Participants in this experiment took a 50- to 55-minute walk in a park setting or in a downtown setting (the setting in which participants first walked [park or downtown] was counterbalanced). Cognitive and mood measures were taken before and after the walk. The cognitive measure was a digit span task in which participants heard a set of digits and then had to recall as many of the digits as they could in reverse order. The mood measure was obtained with a standardized assessment of mood.

Let's analyze the design of this experiment. The two within-subjects factors were the place where the walk was taken (park or downtown environment) and when the cognitive and mood measures were obtained (before and after the walk). The order in which walks were taken (park or downtown) was counterbalanced to test for carryover effects.

The results showed that after walking in the park (nature) participants recalled more digits (M = 1.5) than after walking in the downtown environment (M = .5). There was no effect of order nor were there any interactions with order. The authors concluded that there were greater cognitive benefits from walking in nature than from walking in an urban environment.

QUESTIONS TO PONDER

1. What is a factorial design?
2. What are the advantages to using a multifactor experimental design?
3. What is a main effect?
4. What is an interaction, and how does it differ from main effects?

Higher-Order Factorial Designs

You can extend the simple two-factor, four-cell design to include any number of levels of a given factor and any number of factors. These are **higher-order factorial designs**. However, practical considerations limit the usefulness of these designs if you try to extend them too far. Two important problems concern the number of subjects required for the design and the complexity of potential interactions.

You should probably use at least five subjects per group for a reasonable ability to detect the effects of the independent variables. The number of required cells can be calculated by multiplying together the number of levels of each factor in the design. In a two-factor experiment with two levels of each independent variable, there would be four (2 × 2) cells. If you were to conduct this study using 5 participants per group, you would need to recruit 20 participants. Adding a third two-level factor would produce a design that required eight (2 × 2 × 2) groups, or 40 (5 × 8) participants. If each factor had three levels instead of two, you would need 27 (3 × 3 × 3) groups multiplied by 5, or 135 participants! And this estimate uses a *minimum* number of participants per group. It would be preferable to use more participants for statistical reliability, but could you afford the time and money required to run such an experiment? As you can see, extended factorial experiments get out of hand quickly.

The second problem with extended factorial designs concerns the number and complexity of the resulting interactions. With three factors, you get three main effects (one for each factor), three *two-way interactions* (Factor A × Factor B, Factor A × Factor C, and Factor B × Factor C), and a *three-way interaction* (Factor A × Factor B × Factor C). The two-way interactions are each similar to the simple interaction found in the two-factor design. For example, the A × B interaction represents the interaction of Factors A and B, averaged over the levels of Factor C. However, what is the A × B × C interaction? This interaction occurs when the A × B *interaction* changes depending on the level of Factor C! (Other interpretations are also possible.) Adding a fourth variable to your design adds further interactions, including the dreaded A × B × C × D four-way interaction, which few ordinary mortals can comprehend. Because data resulting from such designs are difficult to analyze and interpret (not to mention expensive because of the large subject requirement), most investigators limit factorial designs to no more than three factors.

In designs with three factors, the simple effects observed at each level of one of the factors consist of two-way interactions of the other two factors. You also could examine the simple effects of each combination of levels of two of the factors. Of course, the same logic can be applied to designs having more than three factors.

If you are willing to give up information about some of the higher-order interactions, you can include a relatively large number of factors and levels within factors while keeping the requirement for subjects within reasonable bounds. Describing the logic behind these designs and the analyses appropriate to them is beyond the scope of this book. If you are interested in looking into such designs, see the discussion of fractional replications of factorial designs in Edwards (1985, pp. 243–245) or Chapter 8 in Winer (1971).

OTHER GROUP-BASED DESIGNS

This chapter presents a logical progression from the two-level design through the single-factor multilevel design and finally to the factorial design. This progression might lead you to believe that these are the only ways to conduct experiments. This is far from the truth.

Chapter 4 indicated that you should first develop your research questions and then choose your design. This rule applies when you are deciding how to conduct a between-subjects or within-subjects experiment. Situations occur in which a full factorial design is not the best design to test your hypotheses. After all, in a full factorial design, you examine every possible combination of the levels of your independent variables. In some research situations, this may be neither necessary nor desirable. For example, to address specific questions, you may need to add control groups or treatments to the basic factorial plan. Alternatively, some groups or treatments may not be possible (e.g., you may not be able to combine certain drugs).

Another creative design is the *fractional factorial design*. With this design you do not run a full factorial design with all levels of your independent variables crossed. Instead, you select only those levels of your independent variables that specifically test your hypotheses. This design allows you to evaluate a larger number of independent variables within a single design (Stolle, Robbennolt, Patry, & Penrod, 2002).

Designs such as these that do not follow one of the more standard design formats can help you address special questions or deal with unusual circumstances. However, the data that they provide sometimes require special statistical techniques to interpret properly. This may mean extra work in identifying the appropriate techniques and learning to apply them.

In short, the special versions of these designs may pose special problems for you. Nevertheless, you should make the effort. Choose the design that best addresses the questions that you want to answer. If your design does not address these questions, you may save some effort, but the effort you do make will be wasted.

DESIGNS WITH TWO OR MORE DEPENDENT VARIABLES

Just as it is possible to include more than one independent variable in a design, it is also possible to include more than one dependent variable. Indeed, all the designs discussed thus far could be modified to include multiple dependent variables without changing their essential characters. Designs that include multiple dependent variables are termed *multivariate designs*. Those with single-dependent variables (such as previously discussed) are termed *univariate designs*. However, multivariate designs are not limited to experiments. Correlational research that simultaneously measures three or more variables is also termed *multivariate*. Chapter 15 covers multivariate designs in some detail, including experimental, correlational, and mixed strategies.

QUESTIONS TO PONDER

1. What is a higher-order factorial design?
2. What are the advantages and disadvantages of a higher-order factorial design?
3. What are some of the other group-based designs?
4. How do designs with more than one dependent variable work?

CONFOUNDING AND EXPERIMENTAL DESIGN

Chapter 4 defined a *confounding variable* as one that varies along with your independent variable. The presence of a confounding variable damages the internal validity of your experiment. Consequently, you may not be able to establish a causal relationship between your independent variable (or variables) and your dependent variable. One of the most important aspects of experimental design is to develop an experiment that is free of confounding variables.

Sometimes a source of confounding can be subtle and difficult to detect. Imagine that you are ready to run an experiment on the effect of drugs on discrimination learning. You are using rats as subjects. You order your rats, which are delivered in crates, each housing several animals. You begin to take the rats out of the crates and put them in individual cages in the colony room of your animal research facility. The first rats that you catch go in the first row of cages, and the remaining rats go in the second row. You decide to assign the rats in the first row to your experimental group and those in the second row to your control group. You run your experiment and find that the rats in the experimental group perform better than those in the control group and conclude that the drug improves performance. Is there another possible explanation? Unfortunately, there is. It could be that the rats in the experimental group were better at discrimination learning to start with. How can this be, you ask?

Remember how you assigned rats to their cages? You caught them and put them in two rows of cages. The rats that went in the first row were caught first. These rats were slower and easier to catch than those in the second row. It may be that these rats are more docile and easier to handle. It is possible that the more docile rats in the experimental group learned faster because they experienced less stress in your experiment than those in your control group. This, and not the drug that you administered, may have produced the observed difference. You could have avoided this problem by randomly assigning subjects to treatment conditions so that each subject, regardless of what row it was housed in, had an equal chance of appearing in either of your treatment groups.

Confounding also can occur if you are using human participants in an experiment that will take a long period of time to complete. You should be sure that your experimental treatments are spread out evenly over the entire period of the experiment. That is, be sure not to run all of your participants in one condition at the beginning of a semester and all the participants in another condition at the end. It could be that participants who volunteer for your experiment at the beginning of a semester differ in some important ways from those who volunteer later in a semester. By randomly assigning participants to treatment groups, you can avoid this problem.

Another source of confounding is *experimenter bias*, which we discussed in detail in Chapter 5. For example, if you assigned your participants to groups because you thought that certain participants would perform better in one group than in another, you introduced bias into your experiment as a confounding factor. To avoid this problem it would be best to use a blind or double-blind technique (see Chapter 5).

A major source of confounding occurs when your treatment conditions are not carefully conceived and, as a result, unintended variables are introduced whose values change in lockstep with those of the independent variable. The old "Pepsi Challenge" commercial, designed to test Pepsi against Coca-Cola, provides a classic example of this source of confounding. In the original version of the challenge, cups with Pepsi were always marked with an "M," and cups with Coca-Cola were marked with a "Q." The challenge showed that participants preferred the taste of Pepsi over Coca-Cola. However, when researchers from Coca-Cola tried to replicate the challenge, they found that participants chose the cup marked with an "M" even if both cups contained Coca-Cola (Huck & Sandler, 1979). Thus, in the original Pepsi Challenge, it is unclear whether participants were making their choice based on taste or on the letter used to mark the cup.

Avoiding this source of confounding requires paying attention to detail. In the Pepsi Challenge, the cups could have been left unmarked, or the letters could have been counterbalanced. Yet another approach would have been to have conducted a pilot study to determine whether a preference exists for certain letters over others.

The best way to avoid confounding in an experiment is to plan carefully how your independent variables are to be executed. Ask yourself whether there are potential alternative explanations for any effect that you may find. Is your independent variable the only factor that could affect the value of your dependent variable? Careful evaluation of your experimental design and variables and a good knowledge of the literature in your area will help you avoid confounding. Remember, results from a confounded experiment cannot be rehabilitated and are generally useless. Therefore, take care during the design stage of your experiment to eliminate confounding variables. This will ensure an experiment with the highest level of internal validity.

QUESTIONS TO PONDER

1. How does a confounding variable affect the validity of your results?
2. How can confounding variables be eliminated?

SUMMARY

Experimental designs can be classified as between-subjects, within-subjects, or single-subject designs. Between-subjects designs manipulate the independent variable by administering the different levels of the independent variable to different groups of subjects. Within-subjects designs administer the different levels of the independent variable at different times to a single group of subjects. Single-subject designs manipulate the independent variable as the within-subjects designs do but focus on the performances of single subjects rather than on the average performances of a group of subjects.

A key problem for any experimental design is the problem of error variance. Error variance consists of fluctuations in scores that have nothing to do with the effect of the independent variable. Error variance tends to obscure any causal relationship that may exist between your independent and dependent variables.

Several steps can be taken to reduce error variance. In any type of design, you can hold extraneous variables as constant as possible and manipulate your independent variable more strongly. In between-subjects designs, you also can randomize error variance across groups by assigning subjects to groups at random, thus tending to equalize the effect of error variance on mean performance across treatments. Alternatively, you can match subjects across treatments on characteristics that you believe may strongly affect the dependent variable, which will thus tend to eliminate any average differences in these characteristics across the treatments. In within-subjects designs, you can use the same subjects in each treatment and in effect match each subject with him- or herself and eliminate differences between subjects from the analysis. Whether you have used a randomized groups, matched-groups, or within-subjects design, you can then use inferential statistics to assess the probability with which random error variance by itself (in the absence of any effect of the independent variable) would have produced the observed differences in treatment means. If this probability is small, you can be reasonably confident that your independent variable is effective.

Between-subjects and within-subjects designs can be classified according to the number of levels of a single independent variable (two or more than two), the number of independent variables manipulated (single-factor or multifactor), the way in which subjects are assigned to treatments (random assignment, matching, or same subjects in every treatment), and the number of dependent variables (univariate or multivariate). If an independent variable takes on three or more quantitative values, the manipulation is described as parametric; otherwise, it is said to be nonparametric.

Designs may include multiple control groups or treatments to assess the impact of several potentially confounding factors. These additional conditions can be included in both parametric and nonparametric designs.

When subjects are assigned to groups at random, the design is termed a *randomized-groups design*. Such designs are best when subject characteristics do not contribute greatly to error variance. However, when subject characteristics strongly influence the dependent variable, you can reduce error variance created by these characteristics by using either a matched-groups design or a within-subjects design. This control over error variance can improve your chances of detecting any effects of your independent variable. However, compared with randomized-groups designs, matched-groups designs require the extra steps of testing and matching subjects. Both matched-groups and within-subjects designs use somewhat different inferential statistical tests to evaluate the data and may actually be less sensitive to the effects of the independent variable if matching or the use of the same subjects in each treatment does not succeed in reducing error variance. With matched-groups designs, it may be difficult to find enough matching subjects if the design includes several groups. With within-subjects designs, this problem is avoided, but the use of the same subjects in each treatment condition introduces the possibility of carryover, which occurs when exposure to one treatment condition changes how subjects behave in a subsequent treatment condition. Several methods are available for dealing with carryover, including counterbalancing (exposing subjects to treatments in different orders), taking steps to minimize carryover, and using a design that makes any carryover effect into an independent variable.

Two or more independent variables or factors may be manipulated simultaneously in a single experimental design. If each level of each factor is combined once with each level of every other factor, the design is called a factorial design. Each treatment in a factorial design represents a unique combination of the levels of the independent variables, and all possible combinations are represented. Factorial designs make it possible to assess in one experiment the main effect of each independent variable and any interactions among variables. The number of treatment cells required for a factorial design can be computed by multiplying together the number of levels of each factor manipulated.

Some questions are best addressed using designs other than the standard ones. For example, a basic between-subjects factorial design might be expanded to include control groups in order to make comparisons beyond those involving main effects and interactions. Although the statistical analyses required for these designs may be more difficult to define, do not let this difficulty prevent you from choosing the best design for the questions that you want to address.

Multivariate experimental designs include two or more dependent variables. These designs provide information about the effect of the independent variable on each dependent variable and on a composite dependent variable formed from a weighted combination of the individual dependent variables.

Confounding occurs when the effects of uncontrolled extraneous variables cannot be separated from those of the intended independent variables. Nonrandom assignment (in between-subjects designs), carryover (in within-subjects designs), experimenter bias, and ill-conceived experimental conditions are sources of confounding. Steps such as random assignment, blind techniques, and careful assessment of experimental conditions and of potential alternative explanations should be taken to avoid potential confounding.

KEY TERMS

between-subjects design
within-subjects design
single-subject design
error variance
randomized two-group design
parametric design
nonparametric design
multiple control group design
matched-groups design
matched-pairs design
carryover effect
counterbalancing
factorial design
main effect
interaction
simple main effect
higher-order factorial design

11 CHAPTER

CHAPTER OUTLINE

Combining Between-Subjects and Within-Subjects Designs

- The Mixed Design
- The Nested Design

Combining Experimental and Correlational Designs

- Including a Covariate in Your Experimental Design
- Including Quasi-Independent Variables in an Experiment
- An Example of a Combined Design: Is Coffee a Physical or Psychological Stimulant?

Quasi-Experimental Designs

- Time Series Designs
- Equivalent Time Samples Design
- Advantages and Disadvantages of Quasi Experiments
- Nonequivalent Control Group Design

Pretest—Posttest Designs

- Problems With the Pretest–Posttest Design
- The Solomon Four-Group Design
- Eliminating the Pretest

Developmental Designs

- The Cross-Sectional Design
- The Longitudinal Design
- The Cohort-Sequential Design

Summary

Key Terms

Using Specialized Research Designs

In Chapters 8 through 10, we introduced you to a variety of research designs appropriate for nonexperimental and experimental research. Although these designs cover a wide range of conventional research situations, some research questions can be adequately addressed only by using a specialized design. In this chapter, we describe several designs in this category. These include combined between-subjects and within-subjects designs, combined experimental and correlational designs, pretest–posttest designs, quasi-experimental designs, and developmental designs.

COMBINING BETWEEN-SUBJECTS AND WITHIN-SUBJECTS DESIGNS

In Chapter 10, we introduced you to between-subjects and within-subjects designs and described each type's advantages and disadvantages. Both types of design come in versions that allow you to assess simultaneously the effects of two or more independent variables. However, it is also possible (and at times desirable) to combine between-subjects and within-subjects manipulations in a single experiment. This section discusses two designs that combine between-subjects and within-subjects manipulations.

The Mixed Design

A **mixed design** (sometimes called a *split-plot design*) is one that includes a between-subjects factor and a within-subjects factor. The term comes from agricultural research in which the design was first developed (it referred to a plot of land). In the split-plot design, a field was divided into several plots. Different plots received different levels of a given treatment (different pesticides). Each plot was then split into subplots, and each subplot received a different level of a second treatment (e.g., different fertilizers). Thus, each plot received all the levels of fertilizer, but only one level of pesticide. In psychological research, each "plot" is a group of subjects who all receive the same level of the between-subjects variable. Within a given plot, the subplots represent the different levels of the within-subjects variable to which all members of that group are exposed.

A mixed design allows you to assess the effects of variables that, because of irreversible effects or carryover, cannot be manipulated effectively within subjects. (These variables are manipulated between subjects.) A mixed design maintains the advantages of the within-subjects design for the remaining variables.

An Example of a Mixed Design: Does Interpersonal Contact Reduce Stereotypes? An experiment by Lindsay Cameron and Adam Rutland (2006) is an example of a study using a mixed design. Cameron and Rutland were interested in determining whether contact between disabled and nondisabled children would reduce stereotyping of disabled children. It has long been known in social psychology that direct intergroup contact can be effective in reducing stereotypes and prejudice. Cameron and Rutland investigated the effects of indirect "extended" contact on stereotyping of and prejudice toward disabled children.

Cameron and Rutland (2006) randomly assigned nondisabled British schoolchildren between 5 and 10 years of age to three extended-contact groups that were exposed to three different experimental conditions. All three groups heard a story depicting a friendship between a disabled child and a nondisabled child. In the neutral condition, the story minimized the categorization of the friends as disabled or nondisabled and provided no additional information about the characteristics of the two children. In the decategorization condition, the story included information about the children's individual characteristics (e.g., they were kind, liked chocolate, and enjoyed playing computer games). As in the neutral condition, the story minimized the categorization of the children as disabled or nondisabled. In the intergroup condition, the story emphasized that the disabled and the nondisabled child depicted in the story were typical of members of those groups. In all other ways, the story was identical to the version read in the decategorization condition. All participants were interviewed 1 week prior to the story intervention and again 1 week after the story intervention. In these interviews, the participants' attitudes toward disabled children and their willingness to form a friendship with a disabled child were measured.

Before we get to the results of this experiment, can you identify the between-subjects and within-subjects factors in this experiment? Figure 11-1 depicts the mixed design used by Cameron and Rutland (2006). The between-subjects factor is the extended-contact manipulation. Different participants were randomly assigned to each of the three extended-contact groups. The within-subjects factor is the time of measurement (before and after extended contact). All participants were assessed before the extended-contact manipulation and again after the manipulation.

The results of Cameron and Rutland's (2006) study showed a main effect of time of interview. Participants showed significantly more favorable attitudes toward the disabled after the intervention ($M = 4.38$) than before ($M = 3.24$). There was also an interaction between extended contact and time of the interview. This interaction is shown graphically in Figure 11-2. Cameron and Rutland report that there was no significant difference between the first interview and the second interview

	Neutral	Decategorization	Intergroup
Preintervention interview			
Postintervention interview			

FIGURE 11-1 Design of Cameron and Rutland's (2006) mixed experiment. Extended contact condition (neutral, decategorization or intergroup) was the between-subjects factor, and time of interview (preintervention and postintervention) was the within-subjects factor.

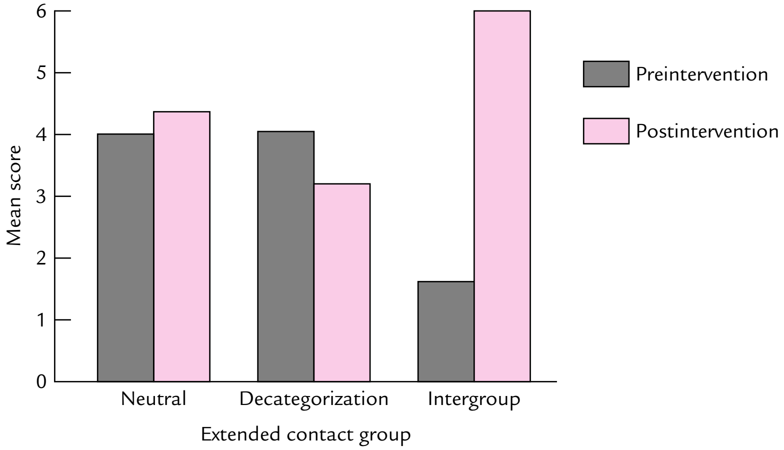


FIGURE 11-2 Bar chart showing an interaction between the extended-contact and time-of-measurement variables.
SOURCE: Based on data from Cameron and Rutland, 2006.

for the neutral and decategorization groups. A significant effect of time of interview was found for the intergroup extended-contact group. As shown in Figure 11-2, attitudes toward the disabled were markedly more favorable after the intergroup extended contact than before it. Cameron and Rutland (2006) conclude that extended contact can be an effective tool in reducing children's prejudice directed toward disabled children.

The Nested Design

Another design that combines within-subjects and between-subjects components is the nested design. In a **nested design**, different levels of a within-subjects independent variable are included under each level of a between-subjects factor. Figure 11-3 shows an example of a nested design. This example includes three levels of a between-subjects factor (A_1, A_2, and A_3). Under each of the levels of A are "nested" three levels of B. Notice that the levels of B found under different levels of A are *not* the same. For example, B_1, B_2, and B_3 appear under A_1 whereas B_4, B_5, and B_6

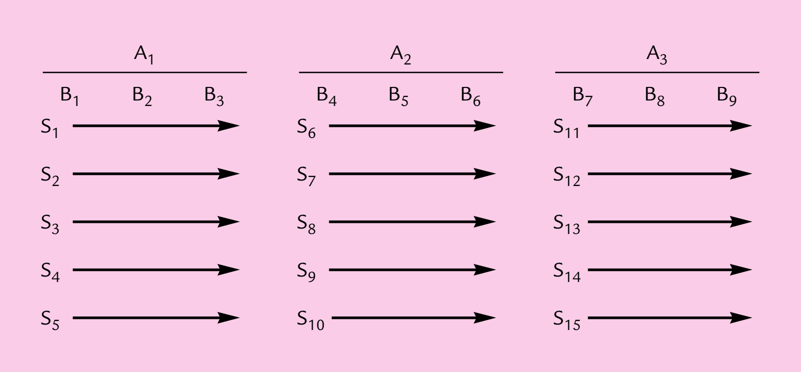


FIGURE 11-3 A nested design that has a within-subjects component.

appear under A_2. Each level of Factor A thus includes a within-subjects manipulation of Factor B although the levels of B included in the manipulation differ with the A level.

Nested designs are more economical than factorial designs in which each level of a factor is completely crossed with every level of every other factor. On the negative side, they do not yield as much information as factorial designs in that you cannot evaluate certain interactions. However, there are situations in which a factorial design is simply not practical. In such cases, a nested design may give you the information you want. Here we discuss two types of nesting: nesting tasks and nesting groups of subjects.

Nesting Tasks Nested designs are useful when you want to include more than one task under a level of an independent variable. For example, imagine that you are conducting an experiment on the effect of a safety campaign on the number of worker injuries. You want to establish that any positive effect of the campaign is not specific to one industry, so you include several companies in your study, representing different industries. However, each company has a different set of jobs. Company X provides jobs A, B, and C; company Y provides jobs D, E, and F; and company Z provides jobs G, H, and I. Thus, jobs are nested under the different companies. Figure 11-4 shows this design.

The major advantage of a nested design like the one shown in Figure 11-4 is that you increase the generality of your results. By demonstrating the effect of the safety campaign across many types of job within each company, you can be more certain that your effect is not limited to a particular type of job. However, the influences of job type and company (e.g., differences in corporate climate) on the effectiveness of the safety campaign are to some extent confounded in this design because the jobs nested under different companies are not strictly comparable.

Nesting Groups of Subjects A nested design also can be useful when you must test subjects in large groups rather than individually. For example, you may find it necessary to test participants during their regularly scheduled class hours. If you tested three classes under each of your experimental conditions, you would then have three

Company X	Company Y	Company Z
Job_A	Job_D	Job_G
Job_B	Job_E	Job_H
Job_C	Job_F	Job_I

FIGURE 11-4 A design with three jobs nested under each company. The jobs nested under each company are different.

	Level of independent variable (A)					
	A_1			A_2		
Nested classes	Class_1	Class_2	Class_3	Class_4	Class_5	Class_6

FIGURE 11-5 Groups of subjects nested under levels of the independent variable.

classes nested under each level of your independent variable. Figure 11-5 illustrates this situation.

In this design, the classes are treated as the subjects rather than as the participants themselves. These classes would be randomly assigned to the experimental conditions, just as you would randomly assign participants to the experimental conditions in a between-subjects experiment. Random assignment tends to average out any differences between classes, making the experimental groups more nearly equivalent.

Nesting groups of subjects within levels of the independent variable should not be done if you can nest only one group under each level of your independent variable. In this case, your experiment is hopelessly confounded because you have no way of knowing whether the differences between groups of subjects across treatment levels occur because of your independent variable or because of something relating to the nested groups (Keppel, 1982).

Nesting several independent groups under each level of your independent variable is analogous to running individual subjects. When you randomly assign individual subjects to conditions in a between-subjects design, you are essentially nesting subjects within treatments. Each group in a nested design can be viewed as a subject nested under a level of your independent variable. As long as you randomly assign groups of subjects to experimental conditions, nesting groups of subjects is legitimate.

QUESTIONS TO PONDER

1. What is a mixed design, and when is it used?
2. What is a nested design, and when is it used?
3. What are the various types of nesting that can be done? Why would you use each?

COMBINING EXPERIMENTAL AND CORRELATIONAL DESIGNS

Experimental designs have the strong advantage of allowing you to not only identify whether relationships exist between variables but also determine whether the relationships identified are causal ones. The strategy requires that you be able to manipulate the suspected causal variable (the independent variable), hold constant as many extraneous variables as possible, and randomize the effects of any remaining extraneous variables across treatments. Unfortunately, holding variables constant can reduce the generality of your findings (using only males as participants, for example, may yield results that do not generalize to females), whereas randomizing their effects across treatments can produce error variance that obscures the effects of your independent variables.

Fortunately, you often can deal with such problems effectively by using a design that combines experimental and correlational variables. In this section, we explore two ways to combine experimental and correlational variables: including a correlational variable as a covariate in an experimental design and including a quasi-independent variable.

Including a Covariate in Your Experimental Design

When participants differ on some variable (such as IQ or reaction time), you can statistically control the effects of this variable by measuring the value of the variable for each participant along with the value of the dependent variable. This additional, correlational variable is called a **covariate**. The name derives from the fact that you expect the covariate to *covary* with the dependent variable. This will be the case if the covariate correlates directly with the dependent variable or with some unmeasured variable that correlates with the dependent variable.

By including a covariate in your experimental design, you can effectively "subtract out" the influence of the covariate (or any variable correlated with it) from the dependent variable. In this way you reduce error variance and improve the sensitivity of your experiment to the effect of your independent variable.

These designs are relatively easy to implement. Simply by collecting additional data on potentially relevant correlational variables, you can convert a standard experimental design into one that examines the impact of those variables on the relationship between the study's independent and dependent variables. For example, in an experiment on jury decision making, you might suspect that a participant's views on the death penalty might be related to his or her willingness to convict a defendant in a case not involving the death penalty. You could measure your participants' attitudes toward the death penalty and then use that measure statistically as a covariate when analyzing your data.

Covariates typically take the form of continuous variables or of discrete variables having a relatively large number of levels. If the covariate in question is discrete and has relatively few levels, it may make more sense to treat it as a quasi-independent variable, as described next.

Including Quasi-Independent Variables in an Experiment

A **quasi-independent variable** is a correlational variable that resembles an independent variable in an experiment. It is created by assigning subjects to groups according to some characteristic that they possess (such as age, gender, or IQ), rather than using random assignment. For example, you might be interested in comparing the effectiveness of two types of incentive programs on sales by retail store clerks. You could include gender as a quasi-independent variable to determine whether male and female clerks tend to respond differently to the two programs.

Because subjects come into the experiment already assigned to their treatment levels, it is always possible that any relationship discovered may be due to the action of some third, unmeasured variable that happens to correlate well with the quasi-independent variable. Even so, the knowledge that a relationship exists may be important. This is especially true when quasi-experimental variables are added to an experimental design.

Such combinations of both experimental and quasi-experimental variables often resemble the factorial designs described in Chapter 10. These combinations yield a main effect for each independent variable, a main effect for each quasi-independent variable, and one or more interactions. The interactions in such designs can be especially illuminating.

An Example of a Combined Design: Is Coffee a Physical or Psychological Stimulant?

The morning ritual is played in millions of homes across the world. The alarm clock rings, and you get out of bed and stumble bleary-eyed into the bathroom. You wash your face, take a shower, get dressed, and then head to the kitchen for your morning pick-me-up: a good strong cup of coffee. For many, the coffee serves as a quick way to get a jolt of energy before beginning the day. We all know that regular coffee contains caffeine, which is a stimulant. Theoretically, it should make us feel better and clear our heads in the morning.

One question that always crops up when discussing the effects of ingredients such as caffeine is whether these effects are due to a physical, pharmacological effect of the ingredient or to expectations about its effect. For caffeine, is it the physically stimulating effect of caffeine that provides the pick-me-up, or is it our belief that caffeine will stimulate us? An experiment conducted by Adam Oei and Laurence Hartley (2005) addressed this question.

Oei and Hartley (2005) approached this question experimentally by using a design that combined one correlational variable and two experimental variables. The correlational variable was whether participants expected that caffeine would stimulate them or believed that caffeine would not stimulate them. Having participants answer a question concerning their expectations about the effects of caffeine created this variable. Participants were *not* randomly assigned to levels of this variable. The two experimental variables were the type of beverage that participants were given (caffeinated or decaffeinated coffee) and the information given about the caffeine content of the beverage that they consumed (participants were told either that the coffee was caffeinated or that the coffee was decaffeinated). The experimental variables were both within-subjects variables and counterbalanced using a Latin square design (see Chapter 10).

Participants came to a laboratory where the coffee was made right in front of them. The experimenters had four jars of coffee, two labeled "caffeinated coffee" and two labeled "decaffeinated coffee." One jar of each type was correctly labeled (e.g., caffeinated coffee in the caffeinated coffee jar), and the other was mislabeled (e.g., caffeinated coffee in the decaffeinated jar). This allowed Oei and Hartley (2005) to completely cross the beverage type with the information provided about the beverage type. Figure 11-6 shows the design of the experiment.

One hour after drinking the coffee, participants completed several tasks, including a signal-detection task. In this task, participants were required to indicate across a series of trials whether or not an "x" appeared against a field of visual noise (dots) on a computer screen. One dependent measure was the number of correct detections, or "hits" (saying that an "x" was present when it was actually present).

The results for this measure showed that participants scored more hits after drinking caffeinated coffee than after drinking decaffeinated coffee. (This effect occurred whether or not the coffee was correctly labeled as caffeinated or decaffeinated.) Almost identical results were obtained if participants merely *believed* they were drinking either the caffeinated or decaffeinated coffee, no matter which type of coffee they actually drank. However, in both cases, this effect was reliable only for those who believed that drinking a caffeinated beverage would affect them. Thus, there is support for the idea that both your expectation about the effects of caffeine and the physical effects of caffeine affect your performance. (Alternatively, it could mean that subjects have learned from prior experience how caffeine actually affects them physiologically.)

If you ignore the fact that participant expectation about caffeine is a correlational variable, then Oei and Hartley's (2005) design looks exactly like a 2 × 2 × 2 mixed design (with participant expectancy as the between-subjects factor). This similarity extends even further as the results of the study would be analyzed exactly as those from the equivalent mixed factorial design as well. What differs is the interpretation.

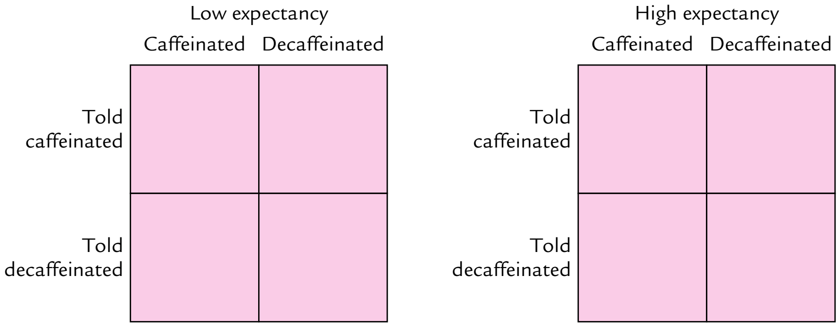


FIGURE 11-6 Design of the Oei and Hartley (2005) combined experiment. The quasi-independent variable was participant expectancy (low or high) and the true independent variables were the actual beverage provided (caffeinated or decaffeinated) and the information provided about the beverage (told caffeinated or told decaffeinated).

Any significant effect of the *experimental* variables in the combined design can be interpreted to mean that the independent variable *caused* changes in the dependent variable. For example, Oei and Hartley (2005) found a main effect of beverage type on the number of hits made in the signal-detection task. More hits were made when caffeinated coffee was consumed than when decaffeinated coffee was consumed. You would be justified in concluding that the presence of caffeine *caused* an increase in hits. Any significant effect of the *correlational* variable in the study *cannot* be legitimately interpreted to indicate a causal relationship, no matter how tempting it may be. Any correlation between two variables could result from the effect of a third, unmeasured variable that influences both variables. A significant effect of participant expectation on performance on the signal-detection task would indicate only that participant expectation about the effects of caffeine is related to performance on the signal-detection task but is not necessarily *causally* related.

Advantages of Including a Quasi-Independent Variable Experimental designs that include a quasi-independent variable allow you to test the generality of your findings across the levels of the quasi-independent variable. In the example, Oei and Hartley (2005) were able to show that the effects of consuming caffeine affected performance on the signal-detection task that did not generalize from participants who believed caffeine would stimulate them to those who did not hold this belief. These results also illustrate a second advantage of including a quasi-independent variable in your design. Had Oei and Hartley analyzed the data without regard to the participants' expectations, these two effects might have canceled each other out, leading to the false conclusion that beverage type had no effect on performance. By including a quasi-independent variable in their design, Oei and Hartley were able to reduce error variance by segregating the data into groups of participants who responded in a similar fashion to the manipulation. In this way, the effect of the independent variable was made clearly visible.

Disadvantages of Including a Quasi-Independent Variable The main disadvantage of including a quasi-independent variable in your design is that the results are frequently misinterpreted. Although paying lip service to the dictum "cause cannot be inferred from correlation," researchers too often discuss their results as if they had established causal links between their quasi-independent and dependent variables. This mistake is encouraged by the fact that the correlational variables sometimes look exactly like experimental variables in the statistical analysis of the data. In Oei and Hartley's (2005) experiment, you may wish to conclude that expecting caffeine to be a stimulant causes an increase in performance when a caffeinated beverage is consumed but does not for those who do not hold this belief. However, one could argue that feeling (or not feeling) stimulated by caffeine causes a person to believe that caffeine will be (or will not be) a stimulant. Because expectation was not experimentally manipulated, the causal status of this variable remains ambiguous.

Another disadvantage of these designs (although a minor one) is the extra effort sometimes required to obtain subjects differing in the required characteristics. Some quasi-experimental variables (such as anxiety level or IQ) require administration of a questionnaire or test before subjects can be classified. Adding quasi-experimental variables to an experimental design also increases the number of groups of subjects required and adds complexity to the analysis of the data.

QUESTIONS TO PONDER

1. When should you consider using a design combining experimental and correlational variables?
2. What is a covariate, and when would you use one?
3. What is a quasi-independent variable, and when would you use one?
4. What are the advantages and disadvantages of including a quasi-independent variable in your research?

QUASI-EXPERIMENTAL DESIGNS

Pure **quasi-experimental designs** are those that resemble experimental designs but use quasi-independent rather than true independent variables. In this section, we examine several types of quasi-experimental design, including time series designs, the equivalent time samples design, and the nonequivalent control group design.

Time Series Designs

The basic time series design is shown in Figure 11-7. In the **time series design** (Campbell & Stanley, 1963), you make several observations (O) of behavior over time prior to (O_1 to O_4) and immediately after (O_5 to O_8) introducing your treatment. For example, you might measure children's school performance on a weekly basis for several weeks and then introduce a new teaching technique (the treatment). Following the introduction of the new teaching technique, you again measure school performance on a weekly basis. A contrast is then made between preintervention and postintervention performance.

Interrupted Time Series Design A variation on the basic time series design is the **interrupted time series design** in which you chart changes in behavior as a function of some naturally occurring event (such as a natural disaster or the introduction of a new law) rather than manipulate an independent variable. In this design, the naturally occurring event is a quasi-independent variable. As with the other time series designs, you make comparisons of behavior prior to and after your subjects were exposed to the treatment.

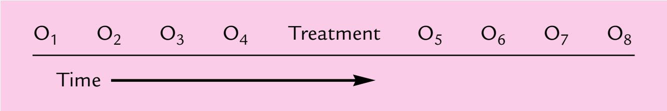


FIGURE 11-7 Basic time series design.

A study conducted by Sally Simpson, Leanna Bouffard, Joel Garner, and Laura Hickman (2006) provides a nice example of an interrupted time series design study. Simpson et al. were interested in whether a change in the domestic violence law in Maryland was related to changes in arrests made by police in cases of domestic violence. Simpson et al. used data collected by the state of Maryland on domestic violence before and after the 1994 change in the law. The researchers focused on statistics collected between 1991 (3 years before the new law went into effect) and 1997 (3 years after the law went into effect). The dependent variable was the likelihood that the police made an arrest when called to a domestic violence scene. Their results showed that the likelihood of police making an arrest increased significantly after the new law was enacted, with a major jump in arrests occurring around the time when the new law took effect. Overall, arrests were made in 27.3% of the cases before the law went into effect and in 39.1% of the cases after the law went into effect.

Basic Data for Time Series Studies In Chapter 8, we defined archival research as research in which you search existing records for your data. Archival data can be used in a time series design. The Simpson et al. (2006) study used arrest statistics from the "Battered Spouse Report" compiled by the state of Maryland. The inclusion of the quasi-independent (enactment of the new law) variable defined the study as a time series study. Hence, in some cases, you may be able to use archival data to investigate possible causal relationships among variables.

You are not limited to archival data when conducting a time series or interrupted time series study. In the example of the impact of a new teaching method on school performance, you could measure ongoing behavior (students' exam scores). In an interrupted time series design, if you know that an event is going to happen (such as the introduction of television to an area in which television is presently unavailable), you can make your observations prior to the introduction of the quasi-independent variable and continue observations afterward.

Equivalent Time Samples Design

A quasi-experimental strategy related to the time series design is the **equivalent time samples design** (Campbell & Stanley, 1963). In this design, the treatment is administered repeatedly. Figure 11-8 shows this design. Note that the treatment is introduced and then observations (O) are made. Next, observations are made without the treatment, followed by a repeat of this sequence. You could repeat the sequence (in any order appropriate to the research question) as many times as necessary. This design is most appropriate when the effects of the treatment are temporary or transient (Campbell & Stanley, 1963).

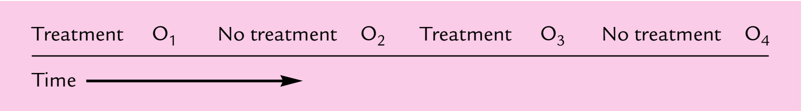

Treatment O_1 No treatment O_2 Treatment O_3 No treatment O_4
Time

FIGURE 11-8 Equivalent time samples design.

Advantages and Disadvantages of Quasi Experiments

One advantage of quasi-experimental designs is that they allow you to evaluate the impact of a quasi-independent variable under naturally occurring conditions. In those cases in which you manipulate the independent variable or even simply take advantage of a naturally occurring event, you may be able to establish clear causal relationships among variables. However, quasi-experimental research does have drawbacks that affect the internal and external validity of your research.

One drawback is that you do not have control over the variables influencing behavior. Another variable that changed along with the variable of interest actually may have caused the observed effect. For example, when the speed limit on the nation's highways was reduced to 55 mph in the 1970s, the accident and death rate noticeably decreased. The temptation is to conclude that driving more slowly caused a reduction in the accident rate.

Although this conclusion is the one that was drawn (and is probably true), other events occurred at the same time. The 55-mph speed limit was instituted during a gasoline shortage. In fact, people drove less and some states instituted "gasless Sundays" on which gasoline was not sold at all. The accident rate could have been reduced because fewer people were on the roads and because those who *were* on the roads drove less after the speed limit reduction than before. Exercise caution when interpreting results from quasi experiments. Be careful to take into account any changes that may have accompanied changes in the variable of interest.

A second drawback to the quasi-experimental strategy also relates to your degree of control over variables. When you are using naturally occurring events as quasi-independent variables, you have little or no control over when the event will occur. For example, you have no control over when a law is changed or a new service is introduced. A research study of any kind requires significant preparation. In the absence of forewarning about an event, you may be caught off guard and not be able to adequately study the behaviors of interest. In the case of a change in a law or introduction of a new service, keeping in touch with current events will provide you with enough advance warning to design a reasonably good quasi experiment. However, in other cases, you may not have such advance warning.

As an example, if you were interested in conducting a prospective study of the impact of Hurricane Katrina on the nearby residents, you probably would have had a problem conducting your research. You would have had to predict when a powerful storm would develop and predict the exact track it would take. Unless you were extremely lucky (or knew more about hurricanes than the experts), you probably would not have been prepared to study the impact of such a natural disaster on human behavior with a quasi experiment. (A more practical design here would make use of available archival data.)

The major problems with the quasi experiment obviously are related to issues of internal validity. Because the researcher does not completely control the quasi-independent variable and other related variables, confounding variables will probably cloud any causal inferences drawn from the data collected. A partial solution to these problems is to include appropriate control groups in your quasi experiment. Campbell and Stanley (1963) suggest some quasi-experimental designs that include such control groups to evaluate the internal validity of your study.

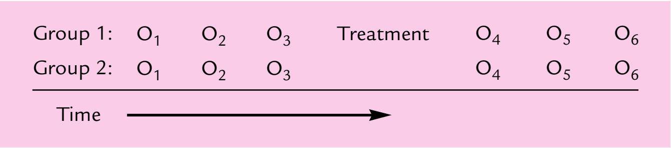
Nonequivalent Control Group Design

In the **nonequivalent control group design**, you include a time series component along with a control group that is not exposed to the treatment. The essence of the nonequivalent control group design is that a comparable group of subjects is chosen and observed for the same period as the group for which the treatment is introduced. The control group is nonequivalent because it comes from a different community. Figure 11-9 illustrates this design.

A study reported by Yun Hee Shin (1999) illustrates a multiple time series design with a nonequivalent control group. Shin investigated the impact of an outdoor walking exercise program on physical health and emotional well-being of elderly Korean women. A sample of women was recruited from an apartment complex for the elderly and enlisted in the walking program. The walking program consisted of a 5-minute warm-up period followed by 30 to 40 minutes of walking, 10 minutes of stretching, and a 5-minute cool-down period. The women in this group constituted the experimental group. A second sample of women recruited from another apartment complex for the elderly made up the control group and was not enlisted in an exercise program (participants in the control group were matched to the participants in the experimental group for age and activity level). Shin measured cardiorespiratory functioning, blood pressure, and resting pulse rate (among several other measures) in both groups before and after instituting the exercise program.

As shown in Figure 11-10, there are no significant differences between the experimental and control groups on three of the measures obtained before the exercise program began. However, as you can see, there are small but statistically significant differences between the groups after the exercise program. In all cases, participants in the experimental group showed better physiological indicators than participants in the control group.

Although the nonequivalent control group design allows you to make comparisons that you ordinarily might not be able to make, there are some drawbacks to the design. First, the validity of the design will be compromised if your two groups differ on some important variable before the study begins (Campbell & Stanley, 1963). For example, living conditions in the apartment complex of participants in Shin's (1999) control group were worse than those in the complex of the experimental participants, and this could account for differences found. To minimize this problem, your groups must be matched as closely as possible prior to your study. Second, if either group is selected on the basis of extreme scores on the pretest, then any shift of scores from pretest to posttest toward the less extreme values may be due to regression toward the mean rather than to the effect of your treatment (Campbell & Stanley, 1963). For example, if the members of the experimental group in Shin's study were selected



Group 1:	O_1	O_2	O_3	Treatment	O_4	O_5	O_6
Group 2:	O_1	O_2	O_3		O_4	O_5	O_6

FIGURE 11-9 Nonequivalent control group design.

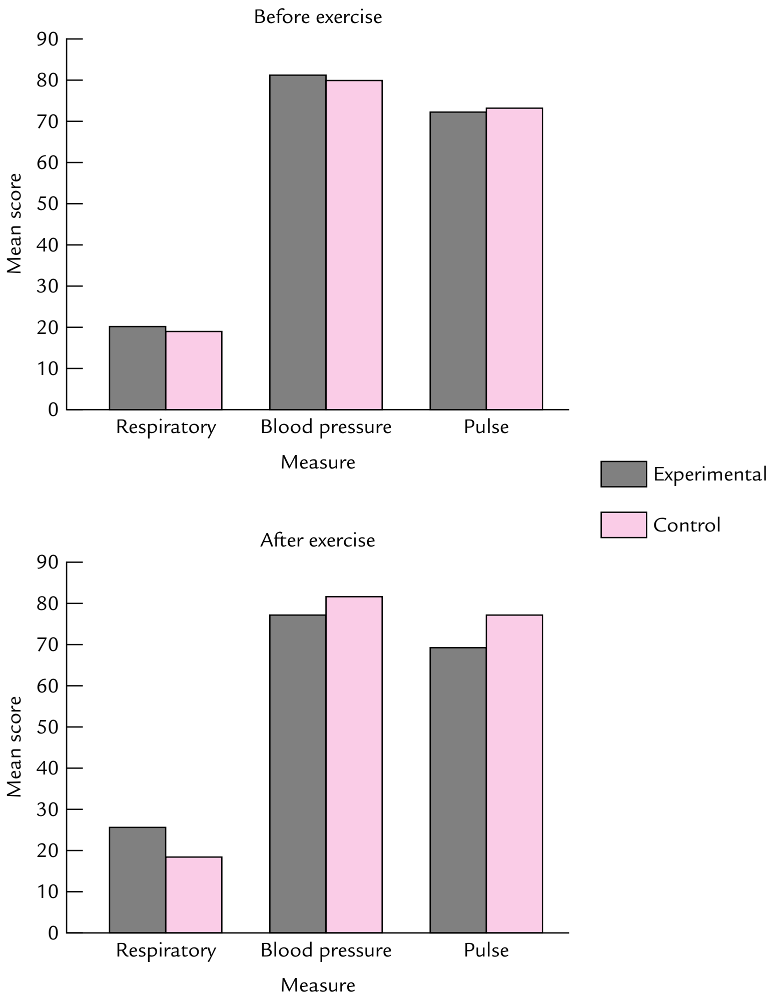


FIGURE 11-10 A nonequivalent control group study of the effects of exercise on the health of elderly women.
SOURCE: Based on data from Shin (1999).

because they were in poor health to start with, any improvement might be due to a tendency for extreme scores to drift toward the average (regression to the mean) and not to the exercise program.

PRETEST–POSTTEST DESIGNS

As the name suggests, a **pretest–posttest design** includes a pretest of participants on a dependent measure before the introduction of a treatment, followed by a posttest after the introduction of the treatment. The pretest–posttest design differs from the previously discussed quasi-experimental strategies in that the pretest–posttest design is a true experimental design (Campbell & Stanley, 1963) that resembles a standard within-subjects design. However, it lacks certain important controls for rival hypotheses.

Pretest–posttest designs are used to evaluate the effects of some change in the environment (including interventions such as drug treatment or psychotherapy) on subsequent performance. You might employ a pretest–posttest design to assess the effect of changes in an educational environment (e.g., introduction of a new teaching method) or in the work environment (e.g., using work teams on an assembly line). The design can also be used to test the effects of an experimental manipulation on behavior. By using a pretest–posttest design, you can compare levels of performance before the introduction of your change to levels of performance after the introduction of the change.

Let's take a look at an experiment that employed a simple pretest–posttest design. Daniel Bernstein, Cara Laney, Erin Morris, and Elizabeth Loftus (2005) conducted an experiment to see if implanting a false memory could affect food aversions. Now, it is well established that false memories of events can be implanted and can be as vivid as real memories. But, can implanting a false memory generate an aversion to foods? This is what Bernstein et al. wanted to find out.

In their experiment, participants completed a food history inventory (FHI) that asked them about their experiences with various foods when they were younger. Participants rated the likelihood that a food-related experience happened to them before age 10 (1 = definitely did not happen, 8 = definitely did happen). Embedded in the FHI were two items concerning getting sick after eating a hard-boiled egg or a dill pickle spear. A week later, participants returned to the laboratory and were given false feedback. They were told, based on their answers to the FHI, that they had gotten sick after eating a particular food. Half of the participants were told that they had gotten sick after eating a hard-boiled egg and the other half after eating a dill pickle. After receiving the false feedback, participants again completed the FHI and a measure of the kinds of foods they were likely to eat or not eat.

When Bernstein et al. (2005) compared participants' responses on the first FHI with those on the second, they found that telling a participant that he or she had gotten sick on a food increased the belief that the event actually happened. That is, participants who were told that they got sick after eating hard-boiled eggs were more confident that this actually occurred than they were before getting the false feedback. The same effect was found for participants who were told that they had gotten sick after eating dill pickle spears. Neither group showed any change to the opposite food item (e.g., those told that they had gotten sick on eggs did not increase their belief that they had gotten sick on pickles). Bernstein et al. also found that the false feedback affected the likelihood that participants would eat the food they thought had made them sick. Those who were told they got sick from eggs were less likely to eat hard-boiled eggs, and those who were told they got sick from dill pickles were less likely to eat dill pickles.

Problems With the Pretest–Posttest Design

Evaluating changes in performance after some change would seem simple: Just measure a behavior, introduce the change, and then measure the behavior again. You should recognize this design as a simple within-subjects experiment with two levels: pretreatment and posttreatment. As with any within-subjects design, carryover effects may confound the effect of the manipulation. Giving your participants the pretest may change the way they perform after you introduce your manipulation—for example, by drawing their attention to the behaviors you are assessing, providing practice on the test, introducing fatigue, and so on. Normally, you would control such carryover effects through counterbalancing. Unfortunately, you cannot counterbalance the pretest and posttest administrations (think about it!). Thus, a simple pretest–posttest design leads to problems with internal validity.

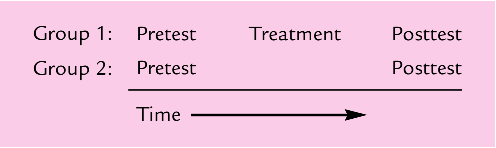
To ensure internal validity, you must include control groups. Campbell and Stanley (1963) discuss pretest–posttest designs extensively. According to Campbell and Stanley, the simplest practical pretest–posttest design should take the form of the diagram shown in Figure 11-11.

As shown in Figure 11-11, the design includes two independent groups of participants. Group 1 (the experimental group) receives your treatment (e.g., false feedback about getting sick after eating a food) between the pretest and posttest. Group 2 (the control group) also receives the pretest and posttest but does not receive the treatment (no false feedback). The pretest and posttest are given to the participants in the experimental and control groups at the same time intervals. Take a look back at the Bernstein et al. (2005) study discussed previously. You will find that there was no true control group in their design. All participants received some form of feedback manipulation.

You may have recognized this design from earlier in the chapter: It is simply a mixed design with pretest–posttest as the within-subjects factor and with treatment versus no treatment as the between-subjects factor. As such, you can use the design to determine the main effect of each factor and the interaction between them.

If the pretest affected the performances of participants on the posttest, you would expect to find a difference between pretest and posttest scores in *both* groups. The same could be said for the effects of any other events, unrelated to your treatment, that might occur between the pretest and posttest administrations: They would be expected to affect the two groups similarly. In contrast, if you found a difference between the pretest and posttest scores in the experimental group only, then you would be justified in concluding that your treatment, and not some other factor, produced the observed changes in performance.

For example, imagine you are interested in whether using computers in a second-grade class affects the children's knowledge of scientific principles. You obtain a sample of 100 second graders. You randomly assign 50 of them to a new teaching method involving computer-aided instruction. These participants constitute your experimental group; the remaining participants are your control group. You give a pretest of scientific principles to all 100 students and then provide computer-aided instruction



Group 1:	Pretest	Treatment	Posttest
Group 2:	Pretest		Posttest

Time →

FIGURE 11-11 The simplest practical pretest-posttest design: A mixed design with pretest-posttest as the within-subjects factor and with treatment versus no treatment as the between-subjects factor.

to your experimental group (the control group continues to get the usual instruction). Finally, you give both groups your posttest.

Imagine that you find that your experimental participants show an average gain of 20 points from pretest to posttest whereas your control group shows an average gain of only 2 points. You could then conclude that your new teaching method and not some other factor was responsible for the observed change. Now imagine that both groups had shown the same 20-point gain. Would you reach the same conclusion? Of course not. Now you would have to conclude that the students showed the same rate of improvement regardless of the teaching method used.

Campbell and Stanley (1963) point out that, for this design to qualify as a true experiment, participants must be randomly assigned to your groups. You *could* use naturally formed groups in a pretest–posttest format, but then the between-subjects component would involve a quasi-independent variable and your conclusions regarding the effect of this variable would be weaker. For example, if you used different classes in your study of computer-aided instruction and found that the class receiving the computers showed a greater increase in performance from pretest to posttest, you could not conclude with any confidence that the computers caused the difference. Perhaps they did, but it is also possible that the teacher of the experimental class simply did a better job of teaching.

Campbell and Stanley (1963) also point out that although the two-group pretest–posttest design ensures a degree of internal validity, it does not preclude potential problems with external validity. Your results may not generalize beyond the immediate research setting. Although this problem can affect any experiment, a particular problem of this design is that participants may be sensitized by the pretest. Having had the pretest, participants may now perform differently than they would have without the pretest. For example, the experimental group may do better than the control group when both groups receive a pretest but not when the pretests are omitted. Campbell and Stanley suggest two remedies to the problem.

The Solomon Four-Group Design

The first remedy is to use a design called the **Solomon four-group design**. This design is illustrated in Figure 11-12. Note that four groups are included in this design. Groups 1 and 2 are identical to those in the two-group design. The additional groups allow you to test for any possible sensitization effects of the pretest. Groups 3 (treatment and then posttest) and 4 (posttest alone) allow you to evaluate the impact of your treatment in the absence of a pretest. By comparing this effect with the impact of your treatment when a pretest is included, you can determine whether inclusion of the pretest alters the effect of your treatment.

A nice example of a Solomon four-group design is provided by a study by Tahira Probst (2003). Probst was interested in studying the relationship between organizational restructuring (e.g., agency mergers and management reorganization) and various measures of employee reactions (e.g., perceived job security and emotional reactions). Participants in the study were employees of a midwestern state whose governor announced a restructuring plan for state agencies. The four groups in

Group 1:	Pretest	Treatment	Posttest
Group 2:	Pretest		Posttest
Group 3:		Treatment	Posttest
Group 4:			Posttest

Time →

FIGURE 11-12 The Soloman four-group design.

Group 1:	Pretest	Affected by reorganization	Posttest
Group 2:	Pretest		Posttest
Group 3:		Affected by reorganization	Posttest
Group 4:			Posttest

Time →

FIGURE 11-13 Design of a study using the Soloman four-group design.
SOURCE: Based on Probst (2003).

Probst's design (as shown in Figure 11-13) were employees affected by the reorganization who were pretested and posttested (the pretest–treatment–posttest group in the Solomon four-group design), employees not affected by the reorganization who were pretested and posttested (the pretest–no treatment–posttest group), affected employees who were posttested but not pretested (the no pretest–treatment–posttest group), and unaffected employees who were posttested but not pretested (the no pretest–no treatment–posttest group). For pretested participants, the pretest was given immediately before the announcement of the reorganization. The posttest was given to all participants 6 months after the announcement. Probst found that the reorganization plan had significant negative effects on employees' perceived job security, commitment to the agency, psychological well-being, and intentions to stay with the agency.

Eliminating the Pretest

Campbell and Stanley's (1963) second remedy to the pretest sensitization problem is to entirely eliminate the pretest. Minus the pretest, this design represents a simple two-group experiment. The decision to eliminate the pretest depends on the question being asked. Situations may exist in which the pretest is needed to answer a research question completely. For example, you may want to know how much material students learn in a given course. However, because some students may come into a course already having some background in the subject, good performance on the final exam would not necessarily mean that the students had learned anything new. To eliminate this problem, you might administer a pretest on the first day of class over the same material to be assessed in the final exam. The results of this pretest would provide a baseline against which you could measure any change attributable to the course.

QUESTIONS TO PONDER

1. What are the characteristics of the time series and equivalent time samples designs?
2. What are the advantages and disadvantages of quasi-experimental designs?
3. How are problems of internal validity addressed in quasi-experimental designs?
4. What is a nonequivalent control group design, and when would you use one?
5. What are the defining characteristics of the pretest–posttest design, and what are the design's strengths and weaknesses?
6. What is the Solomon four-group design, and why would you consider using it?

DEVELOPMENTAL DESIGNS

If you were interested in evaluating changes in behavior that relate to changes in a person's chronological age, you would use one of the specialized *developmental designs* discussed in this section: the cross-sectional design, the longitudinal design, and the cohort-sequential design. These designs represent a special case of quasi-experimental designs wherein a characteristic of the participant (age) serves as a quasi-independent variable. Because age cannot be assigned to participants randomly, it must be used as a purely correlational variable or a quasi-independent variable. Consequently, interpretations that you make from your data should not center on causal relationships between age and behavior change. We also should note that although we are presenting these designs as developmental designs, they often have applications outside of developmental psychology.

The Cross-Sectional Design

Suppose you were interested in evaluating the changes in intelligence with age. One way to approach the problem is to use a cross-sectional design. In the **cross-sectional design**, you select several participants from each of a number of age groups. Figure 11-14 illustrates the general strategy of the cross-sectional design.

In essence, you are creating groups based on the chronological ages of your participants *at the time of the study*. Different participants form each of the age groups. In a cross-sectional study, you do not measure the same participant at different ages.

Assume that you are interested in investigating the developmental changes in intelligence across the life span (birth to death). Your hypothesis is that intelligence increases steadily during childhood and adolescence, levels off during early and middle adulthood, and declines in late adulthood. To evaluate this hypothesis with a cross-sectional design, you would obtain participants representing the different age groups elaborated in your hypothesis. You would then administer a standardized intelligence test (e.g., the Stanford–Binet) to each group and compare results across age groups.

An advantage of the cross-sectional design is that it permits you to obtain useful developmental data in a relatively short period of time. You do not have to follow the same participant for 10 years in order to assess age-related changes in behavior. If you

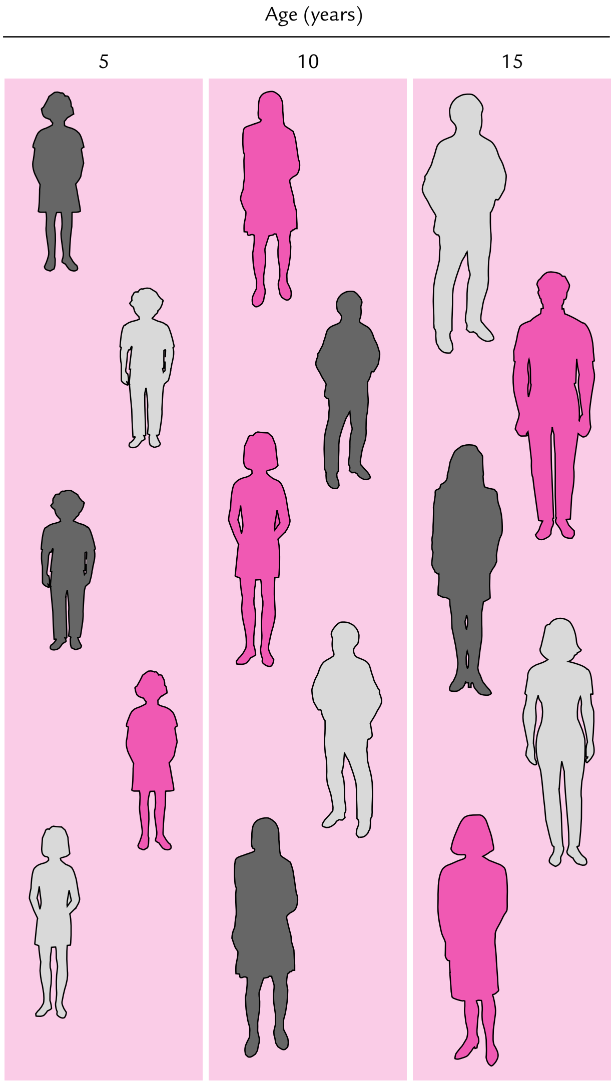


FIGURE 11-14 A cross-sectional developmental design.

found data consistent with your hypothesis, for example, what would you conclude? The purpose of the study was to draw conclusions about changes in intelligence across the life span. The observed decline in intelligence test scores would seem to indicate that intelligence deteriorates with age after middle adulthood.

Yet a serious problem exists with cross-sectional designs that may preclude drawing clear conclusions from the observed differences among intelligence test scores: generation effects. The term *generation effect* refers to the influence of generational differences in experience, which become confounded with the effects of age per se. This confounding threatens the internal validity of cross-sectional studies. Let's say that the participants in your late-adulthood group were 70 years old and participants in your early-adulthood group were 20 years old. Assume that your study was done in 2010. Simple subtraction shows that participants in the different groups not only were of different ages but also were born in different decades: the 70-year-olds in 1940 and the 20-year-olds in 1990.

The fact that participants in different age groups were born in different decades may provide an alternative explanation for the observed differences in intelligence scores, thus threatening the internal validity of your study. The educational opportunities available to those born in 1940 could have differed markedly from those available to participants born in 1990. The observed reduction in intelligence test scores for older participants may be due to poorer educational opportunity rather than to chronological age. Research indicates that the reduced intelligence test scores shown by older participants in cross-sectional studies were indeed caused in part by a generation effect (Anastasi, 1976).

Generation effects are a major problem when you use a cross-sectional design to evaluate age-related changes in behavior of participants of quite disparate ages. The design may be more appropriate when the participants are closer in age. For example, you could use the design to evaluate the changes in the ability to solve verbal problems in children ranging in age from 2 to 6 years. These children would all be of the same generation. In this case, the problem of generation effects may be reduced.

The Longitudinal Design

An alternative to the cross-sectional design, the **longitudinal design**, is illustrated in Figure 11-15. In this design, a single group of participants is followed over some time period. For example, you could obtain a group of participants and give them intelligence tests at 10-year intervals over a 50-year span.

Generation Effects in Longitudinal Designs In one respect, the longitudinal design circumvents the problem of generation effects that plagues cross-sectional designs. Because you are studying people from the same age group, you need not worry about generational effects when drawing conclusions about *that* group. Even so, generation effects may still be of concern in the longitudinal design. Shaffer (1985) points out that longitudinal research has the problem of *cross-generational effects*. That is, the conclusion drawn from the longitudinal study of a particular generation may not apply to another generation.

Suppose a longitudinal study was begun in 1910 and carried out through 1940. Data were collected on attachment (the special bond between parent and child) and other developmental events. Would the conclusions derived from these data apply to the generation that you began to study in 2000? You cannot be sure. Changing attitudes toward child rearing, day care, breastfeeding, and so forth could invalidate the conclusions drawn from data accumulated during the period from 1910 to 1940. Thus, even though the longitudinal design provides important information about developmental trends, you must be careful when attempting to generalize from one generation to another.

Other problems you should consider when choosing a longitudinal design include participant mortality, testing effects, and the time needed to collect even small amounts of data.

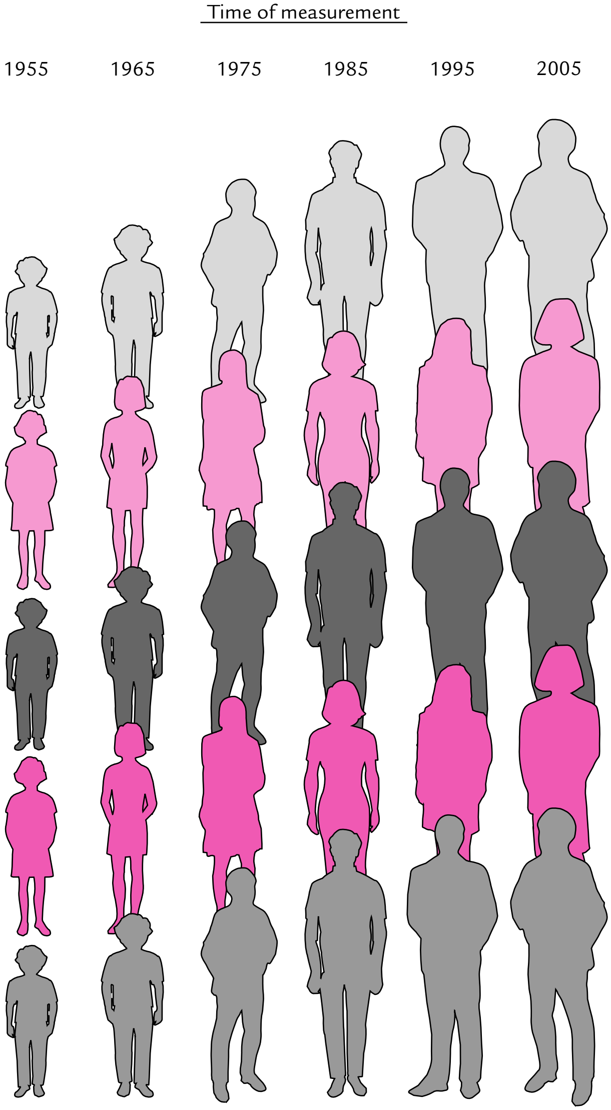


FIGURE 11-15 A longitudinal developmental design.

Subject Mortality The term *subject mortality* refers to the loss of participants from the research. Participants may not complete a longitudinal study because they have moved (and don't notify you of their new address), lost interest in the study, find the study offensive, or have died.

The problem of subject mortality relates directly to the external validity of a longitudinal study. If subject mortality is related to factors such as moving or loss of interest, mortality is less problematic than if it is related to the research. Loss of participants due to factors such as changes in address can be considered random in the sense that moving is as likely to happen to one participant as another. If participants drop out of a study because of the nature of the research (e.g., the methods are stressful or boring), a biased sample results. Those participants who remain in the study (despite finding it offensive) may have special qualities that differentiate them from participants who quit for this reason. Because the loss of those participants biases the sample, the results may not apply to the general population.

Because subject mortality can bias the results of a longitudinal study, you should make every effort to evaluate why participants do not complete the study. Subject mortality may be more problematic with longitudinal research that spans a long period of time. The longer the time period, the more difficult it is to keep track of participants.

Multiple-Observation Effects In a longitudinal design, you make multiple observations of the same participants across time. This very procedure raises problems that may threaten the internal validity of your longitudinal research. Two factors related to multiple observations threaten internal validity.

First, improved performance on the tests over time may be related more to the participants' increasing experience with taking the tests than to changes related to age per se. For example, increases in intelligence scores with age may stem from the fact that participants develop a strategy for taking your test, not from age-related changes in cognitive abilities. Changes in performance due to the effects of repeated testing are referred to as *carryover* (see Chapter 10 for additional information on carryover). The solution to the problem of carryover is relatively simple: (1) Use multiple forms of a test to evaluate behavior at different times, or (2) use different tests that measure the same behavior at different times.

The second problem that results from observing the same participants over time is that other factors tend to arise and become confounded with age (Applebaum & McCall, 1983). For example, if evaluations of behavior are made at 5-year intervals, it is not possible to say conclusively whether the changes observed were due to increased age or to some other factor not related to age. Campbell and Stanley (1963) call this a *history effect* that affects internal validity. An example may help to clarify this point.

Suppose you were interested in evaluating the strength of attachment between a parent and child. You choose a longitudinal design and evaluate attachment behaviors at 2-year intervals. You notice a change in attachment. Is the change caused by the fact that the child has grown older or by other factors such as a shift in attitudes toward children or increased use of day care (Applebaum & McCall, 1983)? It is difficult to know.

The longitudinal design suffers from a problem in which the prevailing attitudes at the time of behavior assessment may influence behavior as much as the change in chronological age. This problem is not as easily handled as is the problem of carryover. You might try to deal with it by including a large enough sample so that any effects of attitudes could be statistically controlled while you are evaluating age-related changes in behavior. However, such large samples of people who are willing to make a long-term commitment to a research project may be hard to find. Once found, those people may constitute a biased sample.

Advantages of the Longitudinal Design Despite its disadvantages, the longitudinal design has an attractive quality. It permits you to see developmental changes clearly. You can witness the development of a behavior. This advantage may make the longitudinal design worth the rather large investment of time it takes to collect data.

Returning for the moment to the issue of changes in intelligence with age, longitudinal research indicates that intelligence for the most part changes very little with age. A few areas of intelligence, such as measures requiring reaction time or perceptual skills, do seem to decline with age. However, the large declines seen with the cross-sectional design do not emerge in the longitudinal data.

The Cohort-Sequential Design

A disadvantage of the cross-sectional and longitudinal designs is their relative inability to determine whether factors other than age are influencing the observed changes in behavior. The **cohort-sequential design**, described by Schaie (1965), combines the two developmental designs and lets you *evaluate* the degree of contribution made by factors such as generation effects. However, the cohort-sequential design does not *eliminate* generation effects. It simply lets you detect them and consider them in interpreting your data.

Figure 11-16 illustrates a cohort-sequential design. Notice that the design embodies the features of both the cross-sectional and longitudinal designs. Along the vertical edge of Figure 11-16 is listed the year of birth. The participants making up one level of this variable (e.g, 1980) constitute a *cohort group*. Specifically, a cohort group consists of participants born at a specified time. In our example, there are three cohort groups: participants born in 1980 (Cohort A), 1990 (Cohort B), or 2000 (Cohort C). These three cohort groups constitute the cross-sectional component when comparisons are made across cohort groups. Along the horizontal edge of the figure is listed the time of measurement. The different measurement times constitute the longitudinal component when we look at a single cohort group across different times of measurement.

By comparing participants from different cohort groups of the same age (e.g., comparing the data from 5-year-olds across cohort groups), we can identify potential generation or cohort group effects. This design is thus useful for evaluating developmental changes in behavior while affording the capability to detect potentially important cohort effects. For a more detailed discussion of this and other developmental research designs, see Applebaum and McCall (1983).

QUESTIONS TO PONDER

1. What are the defining qualities of the cross-sectional developmental design?
2. What are the advantages and disadvantages of the cross-sectional developmental design?
3. What are the defining qualities of the longitudinal developmental design?
4. What are the advantages and disadvantages of the longitudinal developmental design?
5. What is a cohort-sequential design, and when would you use one?
6. What are the advantages and disadvantages of the cohort-sequential developmental design?

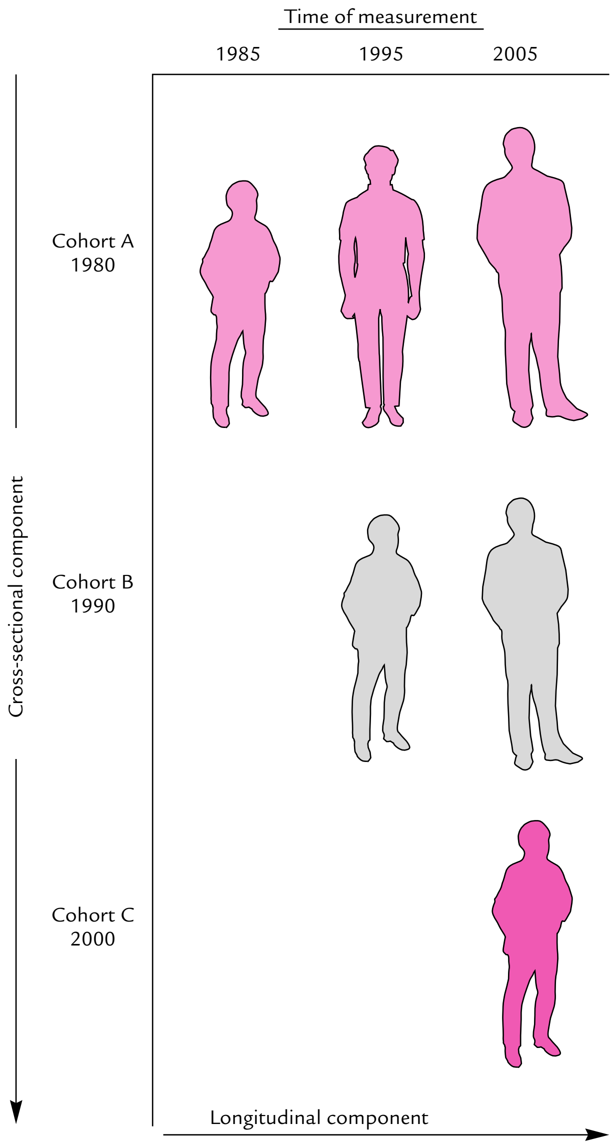


FIGURE 11-16 An example of cohort-sequential developmental design. Comparisons across time of measurement represent the longitudinal component, and comparisons across cohort groups represent the cross-sectional component of the design.

SUMMARY

In cases in which a conventional nonexperimental or experimental design does not meet your research needs, you can use one of the various specialized designs available. The designs discussed in this chapter provide alternatives to conventional designs for specialized research situations.

The mixed design, also known as the split-plot design, combines a between-subjects and within-subjects design. This design allows you to assess the effects of variables that cannot be manipulated in a within-subjects manner due to irreversible carryover effects.

The nested design also combines the between-subjects and within-subjects designs. There are two ways to implement a nested design. First, you can nest tasks under different levels of a between-subjects variable. This is useful when you want to include more than one task under a level of an independent variable (e.g., different lists of words to memorize). Second, you can nest groups of subjects under levels of a between-subjects independent variable. This is useful when you need to run subjects in large groups.

In some research situations, it is desirable to evaluate correlational variables and experimental variables at the same time. There are two ways in which you can add a correlational variable to an experimental design. First, you can include the correlational variable as a covariate. It is measured along with the dependent variable and then used to statistically "subtract out" the effect of the covariate from the dependent variable. Second, you can add the correlational variable as a quasi-independent variable. This variable resembles a true independent variable, but participants come already assigned to their particular level of the variable rather than being assigned to it at random by the experimenter.

Quasi-experimental designs are useful when true experimental designs do not apply to a research situation. They make use of naturally occurring events (such as a change in law or a disaster) as a quasi-independent variable. You do not randomly assign participants to conditions, and in some cases you may not have appropriate control groups. Typically, the quasi experiment suffers in the areas of both internal and external validity. You might consider using a quasi experiment employing the nonequivalent control group design to combat the problems of internal and external validity. However, you must take care when selecting groups for inclusion in this design so that validity is preserved.

Pretest–posttest designs are used when you want to evaluate the impact of some environmental change, such as a new company policy, or naturally occurring behavior, such as the productivity of the company's employees. Behavior is measured before and after the change. A major problem with this design is that participants may be sensitized by the pretest. The Solomon four-group design, a variation of the pretest–posttest design, allows you to test for pretest sensitization effects.

Developmental designs are correlational designs that examine changes in behavior that occur as a function of maturation and experience. The basic developmental designs are the longitudinal and cross-sectional designs. In a longitudinal study, a group of participants is followed over a period of time (weeks, months, or years). This design allows you to observe subtle changes in behavior but suffers from cross-generational problems, subject mortality, and high cost. In cross-sectional research, you study participants of different ages at the same time. This approach is less costly than the longitudinal design, but conceptual problems arise when a wide range exists between the youngest and oldest participants in your study. Generation effects may be a problem in this case. The cohort-sequential design, which combines elements of the longitudinal and cross-sectional designs, allows you to test for generation effects.

KEY TERMS

mixed design
nested design
covariate
quasi-independent variable
quasi-experimental design
time series design
interrupted time series design
equivalent time samples design
nonequivalent control group design
pretest–posttest design
Solomon four-group design
cross-sectional design
longitudinal design
cohort-sequential design

CHAPTER 12

Using Single-Subject Designs

CHAPTER OUTLINE

A Little History

Baseline, Dynamic, and Discrete Trials Designs

Baseline Designs

An Example Baseline Experiment: Do Rats Prefer Signaled or Unsignaled Shocks?

Issues Surrounding the Use of Baseline Designs

Dealing With Uncontrolled Variability

Determining the Generality of Findings

Dealing With Problem Baselines

Types of Single-Subject Baseline Design

Dynamic Designs

Discrete Trials Designs

Characteristics of the Discrete Trials Design

Analysis of Data from Discrete Trials Designs

Inferential Statistics and Single-Subject Designs

Advantages and Disadvantages of the Single-Subject Approach

Summary

Key Terms

The experimental designs described in the last two chapters require that one or more groups of subjects be exposed to the various treatments of the experiment. The data within each treatment are then averaged. The differences among the means are tested statistically to determine the probability that the observed differences could have arisen by chance through the operation of uncontrolled random factors. If this probability is acceptably low, the investigator concludes that the differences are reliable and attributes these differences to the effect of the independent variable.

This chapter presents a very different approach to conducting experimental research, one that focuses on the behavior of individual subjects. It does not depend on averaging across subjects to control the effects of random factors and therefore can be used with few or even only one subject. For this reason, the approach is often called the *single-subject* or *small-*n approach. If you have trouble with inferential statistics, you will be pleased to learn that this approach generally avoids them. This chapter describes the logic of the single-subject approach, indicates conditions under which the single-subject approach is appropriate or inappropriate, and identifies specific single-subject designs.

A LITTLE HISTORY

A major goal of psychology is to understand human and animal behavior. Understanding a particular behavior means knowing what variables influence the behavior and what functional relationships exist between these variables and the behavior. To be useful to psychologists, this understanding must be applicable to individuals.

This emphasis on developing laws that can be applied to individuals dates back to psychology's beginnings as an experimental discipline in the latter half of the 19th century. The psychophysics of Weber and Fechner, the memory experiments of Ebbinghaus, the investigations of perceptual processes by the early Gestalt psychologists, and Wundt's examinations of "mental chronography," as well as the learning experiments of Thorndike and Pavlov, all focused on the behaviors of individual subjects in an effort to understand psychological processes. An extreme example is provided by Ebbinghaus's research, which employed but a single participant—Ebbinghaus himself.

The pioneering experimentalists managed to identify important psychological phenomena, and the functional relationships they uncovered, by and large, have withstood later scrutiny. This was all accomplished without the benefit of inferential statistics, which had not yet been developed.

From the beginning, these early researchers recognized the problems created by apparently random variations in the behaviors of their subjects. One solution to these problems was to repeat the observations many times under a given set of conditions and then average across observations to provide a stable estimate of the "true" values. Although inferential statistics had not yet been developed, researchers knew that estimates based on means become more stable with increasing numbers of observations.

The focus on individual behavior naturally led investigators to adopt a type of within-subjects approach that differs from that described in Chapter 10. In the traditional within-subjects design outlined there, each subject is exposed once to each level of the independent variable, and then scores are averaged across subjects. The method adopted by the early investigators exposed a single subject repeatedly to the different treatments and then averaged across exposures within each treatment. The result was a functional relationship between independent and dependent variables that applied (strictly speaking) to the one individual from whom the data were collected. Functional relationships from different individuals were then compared to determine the generality of the relationships.

Despite intersubject variability, the approach worked because of three factors. First, a very large number of observations were collected from a single subject, thus allowing momentary fluctuations to average out. Second, to the extent possible, incidental factors that might contribute unwanted variability were rigidly controlled. For example, Ebbinghaus ate the same meal at the same time each day during the years he studied his own memory processes (Fancher, 1979). Third, the investigators focused their attentions on powerful variables whose effects could be detected easily against the remaining background of uncontrolled variability.

Of course, certain problems could not be attacked with this approach. These problems involved treatments that produced irreversible changes in subject behavior or that exerted very weak effects on the dependent variable or contained dependent variables that could not be stabilized through rigid control of experimental conditions. Such problems required an approach that could extract the relatively weak signal of the independent variable from the noisy background of random variation. Inferential statistics were developed for these cases.

The application of statistical techniques to the study of individual differences was pioneered by Sir Frances Galton (a cousin of Charles Darwin) in the late 1800s. The first correlational statistic was developed by Karl Pearson under the guidance of Galton and laid the groundwork for the application of statistical techniques to other problems in psychology.

The next major step in the evolution of the statistical revolution came in the 1920s and 1930s when Sir Ronald Fisher and other statisticians developed the rationale of inferential statistics to provide some of the first statistical tests. Soon researchers in psychology recognized that these statistical techniques provided powerful tools for dealing with uncontrolled variability. Inferential statistics were adopted, and the single-subject approach waned in popularity. By 1950 it was virtually impossible to publish research in a respectable psychological journal unless the data had been subjected to an appropriate statistical test and were judged to be reliable.

Meanwhile, some die-hard researchers persisted in using the old nonstatistical, single-subject approach. Most prominent among these was B. F. Skinner. Focusing his efforts on the effects of environmental stimuli on the motor behavior of rats, Skinner developed a highly controlled laboratory environment to observe and record selected behaviors of his subjects. Electromechanical equipment (such as clocks, relays, and switches) was used to gain precise control of environmental stimuli, to program the experimental contingencies, and to define and record the behavioral responses. Skinner and his students continued the tradition of observing and analyzing the behavior of individual subjects. In the process, they developed several methodological refinements that extended the power and usefulness of the single-subject approach.

Unfortunately for Skinner and his followers, their unwillingness to use inferential statistics to establish the reliability of their findings made it increasingly difficult for them to get their results published. In 1958 they attacked this problem by establishing their own journal, the *Journal of the Experimental Analysis of Behavior* (*JEAB*). Eventually, researchers using the single-subject approach were able to convince others of the validity of the method. Today the approach is widely accepted, and experiments using it are being published in many other psychology journals. Because the method specifically focuses on changes in the behavior of the single subject, it has gained widespread acceptance in applied situations in which it has been used to assess the effectiveness of behavioral change programs and therapies in the treatment of individuals.

In 1968 the publisher of *JEAB* launched a second journal, the *Journal of Applied Behavior Analysis* (*JABA*), to publish single-subject research on applied problems. This research provides empirical support for the effectiveness of behavioral management techniques employed by practitioners of applied behavior analysis. Since then *JABA* has been joined by a number of other journals focusing on behavior analysis, such as *The Behavior Analyst*, which began publishing in 1978.

As this brief review indicates, single-subject designs have a long and respectable history and have emerged again into acceptance after being temporarily eclipsed by group-based designs.

BASELINE, DYNAMIC, AND DISCRETE TRIALS DESIGNS

Although single-subject designs come in a variety of forms, all these forms can be categorized into one of three basic types: baseline designs (developed primarily by B. F. Skinner and his followers), what we will call "dynamic" designs, and discrete trials designs (the type used most often by early researchers). Today when researchers refer to "single-subject designs," they usually mean baseline designs. Dynamic designs, which are closely related to baseline designs, are less common but becoming more popular as researchers focus on understanding the dynamics (moment-by-moment changes over time) of behavior. Discrete trials designs are still in use, especially in areas such as psychophysics in which the emphasis continues to be on the performances of individual subjects.

These three types of single-subject design are sufficiently different from one another to require separate treatment. The next sections describe the logic of baseline designs and indicate how to analyze and interpret the results obtained from such designs. We then focus briefly on dynamic designs. Finally, the last part of the chapter describes the discrete trials approach and examines the issues surrounding the use of statistical techniques with single-subject designs.

BASELINE DESIGNS

The group-based experimental designs that we have discussed in previous chapters depend on averaging to even out across the various treatment conditions the effects of any uncontrolled, extraneous variables on the dependent variable. You perform the experiment and then use inferential statistics to evaluate the significance of any differences in mean performance that do emerge between treatments. The statistical analysis is used to decide whether those differences are *reliable*. If the results are reliable, then you would expect to reproduce essentially the same results if you were to repeat, or *replicate*, the study.

In contrast to group-based designs, a **baseline design** focuses on the behavior of a single subject both within and across the experimental treatments and does not rely on averaging to deal with uncontrolled variability. Within a treatment condition, the behavior of interest is sampled repeatedly over time and plotted to create a **behavioral baseline**. This baseline typically changes over time as the effect of the exposure to the treatment condition develops and also in response to the effects of uncontrolled variables. For example, the baseline may rise for a time and then show no further change except for small, unsystematic fluctuations. The subject typically remains under a given experimental treatment until the baseline meets a **stability criterion**, which imposes an objective rule for deciding that the baseline has stabilized. When the behavior has stabilized in this way, the subject is then exposed to the next treatment condition during which the baseline is again plotted until it becomes stable.

After the subject has been exposed to each experimental treatment, these conditions are then repeated. In the simplest case, the design would involve exposing the subject to two conditions: a **baseline phase**, to assess behavior in the absence of the treatment, and an **intervention phase**, to assess behavior during application of the treatment. The subject would be exposed to each of these phases twice, yielding what is called an **ABAB design** where A and B represent the two phases. This immediate **intrasubject replication** of each phase allows you to establish the reliability of your observations within each phase. To the extent that your observations are reliable, the level of baseline observed under one exposure to a phase will be recovered during reexposure to that same phase. In other words, intrasubject replication helps you establish the *internal validity* of your findings.

TABLE 12-1 Characteristics of the Single-Subject Baseline Design

1. Individual subjects are observed under each of several phases. Multiple observations of a target behavior are recorded in a phase before the next phase begins.
2. Extensive observations are made during the baseline phase to establish a behavioral baseline against which any changes due to the independent variable are compared. A behavioral baseline will also be established during the intervention phase.
3. Each subject is observed under all phases, with each treatment phase repeated at least once. This repetition, or intrasubject replication, establishes the reliability of the findings.
4. Subjects usually remain in each phase until a stability criterion is met.
5. Multiple subjects may be included in the experiment. This intersubject replication helps establish the generality of the findings across subjects.

When you move from baseline to intervention and then back to baseline, this return to a previous phase is termed a **reversal strategy** and is designed to assess whether any changes in baseline level produced by the intervention are reversible. If so, then you should be able to recover the original baseline.

Despite the name "single-subject design," most studies of this kind include more than one subject to provide what is termed **intersubject replication**. The purpose of intersubject replication is to establish the *external validity* of your findings. To the extent that different subjects show similar changes in baseline levels across the experimental conditions, you demonstrate that your effects are not unique to a particular subject.

Table 12-1 summarizes the characteristics of the single-subject baseline design.

An Example Baseline Experiment: Do Rats Prefer Signaled or Unsignaled Shocks?

To illustrate the baseline approach to single-subject design, we describe a typical example, part of a larger experiment that investigated whether rats prefer a schedule of signaled shocks over an equivalent schedule of unsignaled shocks (Badia & Culbertson, 1972). The subjects were tested individually in a small operant conditioning chamber equipped with a response lever, a house light, and a floor of metal rods that could be electrified to deliver a brief shock to the rat's feet. (The shocks are similar to static electric pokes and do not harm the rat.)

In the baseline phase, each subject received a series of "training" sessions to familiarize the rat with the characteristics of each shock schedule and to establish a behavioral baseline. At times the chamber house light was off, and at other times it was on. When the light was off, shocks occurred unpredictably according to a random schedule at an average rate of one shock every two minutes (unsignaled schedule). When the light was on, shocks continued to occur on the same schedule, but each shock was immediately preceded by a 5-second warning tone (signaled schedule). Each session provided equal experience with the two schedules.

During the baseline phase, responses on the lever had no effect on conditions, but the number of responses during each session was recorded. At the end of each session, the percentage of responses out of the total possible was calculated and plotted to provide the behavioral baseline. The investigators continued to train each rat until three successive points on the baseline remained within a 10% range (stability criterion). The subject was then placed in the intervention phase of the experiment.

During the intervention phase, the rat was placed on the unsignaled schedule (identified by darkness). Pressing the lever now "bought" the rat 1 minute of time on the signaled schedule. The house light turned on to indicate that the signaled schedule was now in effect, and any shocks that happened to be programmed during the minute were preceded by the warning tone. When this 1-minute "changeover period" ended, the house light was extinguished, and the unsignaled schedule was automatically reinstated. At this time, the rat could buy another minute in the signaled schedule by again pressing the lever. The number of "changeover responses" on the lever was recorded, and as in the baseline phase, the percentage of responses out of the total possible was calculated and plotted. Intervention-phase sessions continued until the stability criterion (three successive points on the baseline within a 10% range) was again met. The baseline phase was then repeated, followed by a second exposure to the intervention phase, to provide an intrasubject replication of each phase.

Note that the two shock schedules were identical except for the signal and that the rat could neither avoid nor escape the shocks. Would the rats nevertheless press the lever to get into the signaled schedule during the intervention phases? The answer to this question can be found in Figure 12-1, which shows the level of responding during the final three sessions in each phase for each rat.

During the initial baseline phase (during which responses had no programmed consequences), the level of responding of each rat remained low, typically around 10% of the maximum possible rate (see the first panel of Figure 12-1). However, note the dramatic changes that occurred when subjects were placed in the intervention phase: Response rates zoomed up to over 85% (second panel). The subsequent return to baseline conditions produced an equally dramatic effect as response rates fell back to the low levels obtained during the first baseline phase (third panel). Finally, response rates jumped back to high levels when subjects were returned to the intervention phase (fourth panel).

The fact that response rates were high in the intervention phases would mean little without the comparison provided by responding in the baseline phase when responses did not produce the signaled schedule. This comparison shows that response rates were high *only* when responses did produce the signaled schedule.

Badia and Culbertson (1972) concluded from this experiment that rats strongly prefer the signaled shock schedule over an equivalent unsignaled schedule although it was unclear why. Evidently, the signaled schedule contains a powerful source of reinforcement that was capable of generating high rates of responding during the intervention phase. The findings, which could not be explained adequately by known principles of conditioning, led to an extensive series of follow-up studies that sought to clarify the sources of reinforcement and their impact on preference responding (reviewed in Badia, Harsh, & Abbott, 1979).

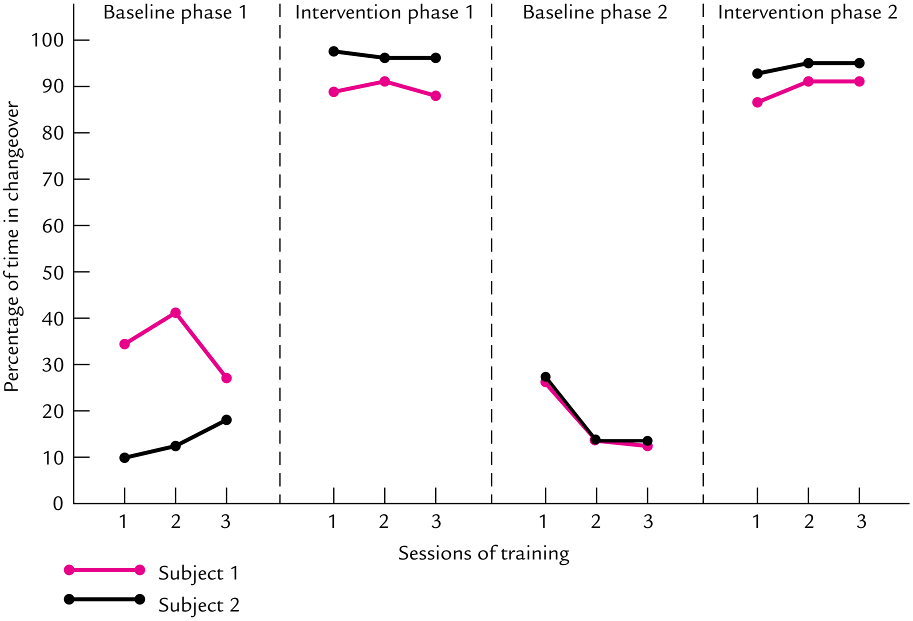


FIGURE 12-1 Results of a single-subject experiment on rat preference for scheduled shock over unscheduled shock.
SOURCE: Data from Badia and Culbertson, 1972, table 3.

QUESTIONS TO PONDER

1. How were single-subject designs used in the early days of behavioral research?
2. What are the major characteristics of the single-subject baseline design?
3. What is a behavioral baseline?
4. Why is it important to establish a behavioral baseline in a single-subject design?
5. What is a stability criterion, and why is it important?
6. What is an ABAB design, and how does it relate to intrasubject replication?
7. What are intrasubject and intersubject replication, and what do they tell you?

Issues Surrounding the Use of Baseline Designs

When using baseline designs, you may have to grapple with a number of issues. These include what stability criterion (if any) to adopt, how to deal with uncontrolled variability (unstable or drifting baselines), and how to cope with irreversible baselines.

Choosing a Stability Criterion Systematic changes such as those due to learning or habituation show up in the behavioral baseline as trends toward increasing or decreasing values. Such trends usually occur immediately after a change to a new phase when behavior is in transition from one stable level to another. Optimally, your stability criterion should guarantee that your subjects will remain in a given phase only until the baseline shows no further systematic changes, and no longer.

Choosing a good stability criterion is something of an art. If your stability criterion is too stringent, your baseline may never achieve it, and you will not be able to proceed to the next phase. Yet if your stability criterion is too lax, you may proceed to the next phase before your subject's performance has actually stabilized. As a result, your "stable" baseline values will not accurately reflect the effect of your independent variable. Developing a good stability criterion may require some pilot work in which you observe your baseline until it shows no long-term trends. You then attempt to identify a stability criterion that would have allowed you to stop sooner without including transitional data. In the example experiment, experience with the percentage measure used indicated that the baseline was likely to remain stable if the data remained within 10% across three successive sessions.

Because only the stable performances truly represent the long-term effect of an independent variable on the dependent variable, you usually report only the data that meet the stability criterion. This was done in the Badia and Culbertson (1972) example: Figure 12-1 omits the transitional data from the plot.

During the experiment, the only way to determine whether your baseline has met the stability criterion is to update your plot after each session and then examine it. If you fail to keep your plot current, you may be shocked to discover that you have run your subject through another session under the previous phase when you should have changed to a new phase. Because this is both a waste of time and a violation of experimental procedure, you can appreciate how important it is to keep the baseline up to date when using a baseline design.

Transitional Behavior and the Stability Criterion By imposing a stability criterion, the single-subject baseline approach removes transitional data (when behavior is changing between stable levels) from the analysis. Of course, if the focus of the experiment is on transitional behavior, then using a stability criterion will not reduce the variability in the data of interest. However, it is still useful because it indicates that the transition is over and you can stop collecting data within that phase.

Stability Criterion Versus Fixed Time or Trials The group approach encourages you to design experiments in which all subjects receive the same amount of exposure to each treatment. If some subjects reach stable levels of performance in the allotted time and others do not, then the data from each treatment will reflect varying mixtures of average transition rates and average levels of steady-state performance. Such differences between treatments may turn out to be statistically significant yet misleading. Employing a stability criterion in such cases helps ensure that the data will reflect terminal levels of performance for all subjects under a given treatment condition.

There are times, however, when it is important for experimental reasons to keep the amount of exposure to a given treatment constant across subjects prior to introducing the next treatment. In that case, you would probably choose to end the phase according to some criterion other than a stable baseline (e.g., after a certain amount of time in the phase). However, because you continue to monitor the baseline of each subject, you can determine how stable each individual's behavior was in the final stages of exposure to the phase.

Melissa Anglesea, Hannah Hoch, and Bridget Taylor (2008) provide an example of the use of an ABAB design in an applied setting. In such settings, the decision to end a phase often depends on factors other than the meeting of an explicit stability criterion, such as time constraints. Anglesea et al. (2008) were concerned with the eating behavior of three teenage boys with autism. Although the boys were capable of feeding themselves, they wolfed down their meals. To get the boys to eat at a normal rate, Anglesea et al. (2008) attached a vibrating pager to each boy's belt and trained each boy (using physical guidance of the boy's hand, verbal reinforcement, and a fading procedure) to place his hand on the pager, wait until the pager vibrated, and then take a bite of a target food. During the baseline phase, the pager was deactivated. During the intervention phase, the pager vibrated every so many seconds, an interval that matched the eating rate of a normal adult. Figure 12-2 shows the total seconds of eating time and the total number of bites taken during the observation periods across the several sessions of each phase. The intervention clearly increased eating time without affecting the amount eaten per bite, as the total number of bites required to consume the food remained essentially constant across phases. Note the success of both the intrasubject and intersubject replications.

Judging Differences in Stable Performance Across Phases Large and consistent differences in performance levels across phases stand out when there is comparatively little variation within each phase, as in the Anglesea et al. (2008) data just presented. Judging whether reliable differences exist is more difficult when baselines are more variable and the difference between their levels is relatively small. Matyas and Greenwood (1990) had students taking a postgraduate course in single-case design and analysis evaluate computer-generated graphs depicting simulated data from an AB design (a baseline phase followed by an intervention phase). The graphs differed in effect size, level of variability, and autocorrelation. False alarm rates (declaring that the intervention was effective when it was not) were "surprisingly high" (16% to 84%) and sensitive to both the level of variability and degree of positive autocorrelation. However, miss rates (declaring that the intervention was not effective when it was) were "relatively low" (0% to 16%). These results suggest that the traditional "eyeball" evaluation may be less conservative with respect to the identification of treatment effects than commonly realized. Similar concerns about the visual evaluation of data from single-subject designs were raised by Kazdin (1978).

Such concerns have led some researchers to propose alternative methods of evaluation. For example, Fisher, Kelley, and Lomas (2003) have developed a "conservative dual criterion" (CDC) method to evaluate differences in trend across phases. Steward, Carr, Brandt, and McHenry (2007) found that this method was superior to the use of lectures to train six university students to visually inspect AB-design graphs. (See Fisher et al., 2003 for a description of the CDC method.)

Other researchers have proposed using statistical methods, such as randomization tests, to discriminate differences in baseline levels, but the suitability of such tests is still controversial. We discuss this issue near the end of this chapter.

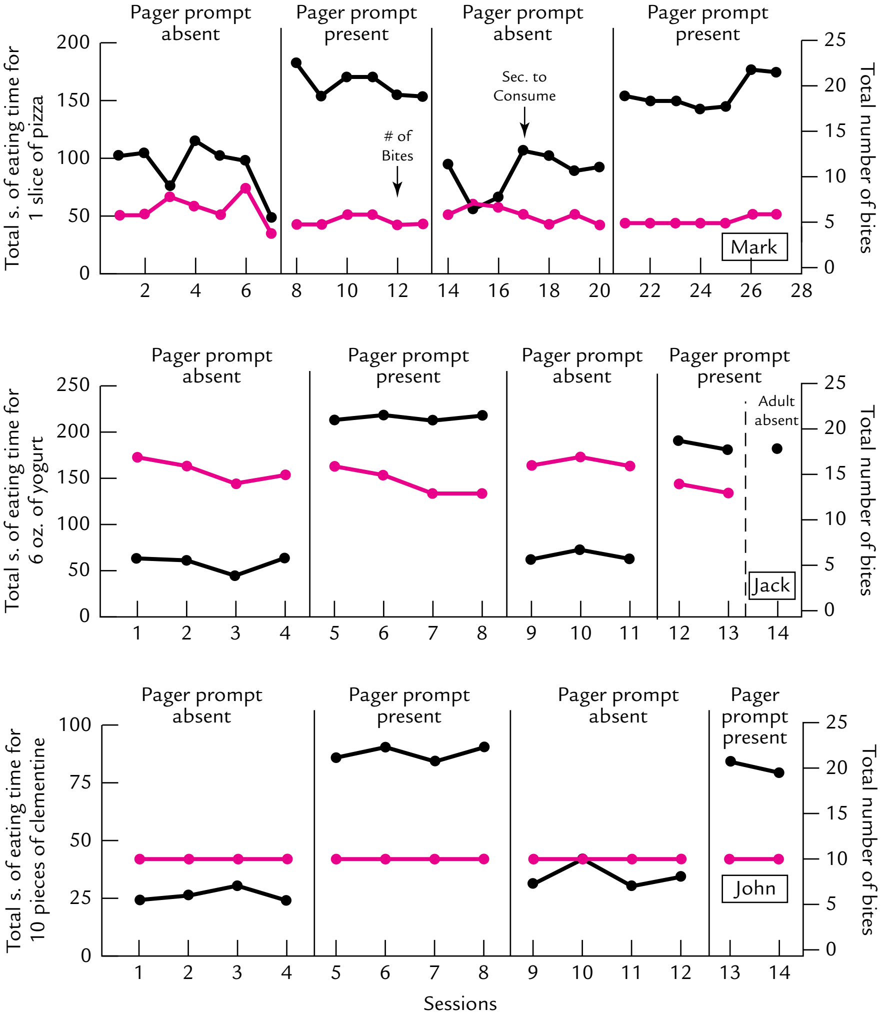


FIGURE 12-2 Results of the eating-behavior study.
SOURCE: Angelesea, Hoch, & Taylor, 2008; Reprinted with permission.

Dealing With Uncontrolled Variability

Within the single-subject approach, extraneous variables produce uncontrolled variability of the baseline within each phase. In the group approach, this variability is handled by averaging data across subjects, but in single-subject designs it is handled instead by tight experimental control. Uncontrolled variability can be reduced only if you can identify its sources. Consequently, the single-subject researcher makes an effort to identify the possible sources of variability. The first step in this process is to graph the data from each subject and look for uncontrolled variability in the baseline, which will be evident when the data points on your graph show moderate to high levels of instability across observation periods.

It is of course unreasonable to expect a given subject to show exactly the same pattern of behavior across observational periods or to expect different subjects to display identical patterns of behavior in a given phase. You must decide how much variation is acceptable. If the observed variation is within acceptable limits, then you (much like the group researcher) consider the observed effects to be reliable. Unlike the group researcher, however, you may still be concerned with uncontrolled variation, despite the emergence of a clear relationship between the independent and dependent variables. In your next experiment, you would then take steps to bring this variation under control.

The difference between the single-subject and group approach is a philosophical one. The group approach assumes that if experimental controls fail to reduce uncontrolled variation, then statistical methods should be used to control it. The single-subject approach assumes that if experimental controls fail to reduce uncontrolled variation, then one should endeavor to identify the extraneous variables responsible for it and bring them under experimental control.

When extraneous variables contribute strongly to variation in the dependent variable, identifying them should help you to better understand the behavior in question. The single-subject approach strongly encourages you to identify these important sources of behavioral influence. The group approach does not do this because the effects of the sources are hidden from view during the averaging process.

Of course, data collected across replications of a given phase usually will be similar, not identical. How similar do they have to be, and in what ways, before you can say that the results have been replicated? The answer depends on the degree of control you have over the dependent variable within a treatment condition (its stability) and on the questions that the experiment was designed to answer.

If you have a relatively high degree of control over your dependent variable within a phase, variation in the baseline across successive observations will be minimal. Any effect of the independent variable will be clearly visible as a shift in performance upward or downward relative to this baseline. If each replication of a given condition produces levels of performance that overlap the levels observed in previous administrations of the same conditions, then the reliability of the data is unquestionable.

If your degree of control over the dependent variable is relatively low, the baseline will be variable and the effect of the independent variable will be more difficult to detect. Variation in baseline levels may occur both within and between replications of the same conditions. Variations between replications can occur both as a result of chance (the data points are varying and happen to be higher or lower during replication) and as a result of carryover. Nevertheless, changes in behavior induced by a particular treatment may be consistent in direction and approximate size.

Figure 12-3 shows an example of such a case. The graph indicates the percentage of study behavior during class for Robbie, a disruptive third grader (Hall, Lund, & Jackson, 1968). During the baseline phase, Robbie's study-behavior varied a fair amount but never exceeded 45%. When the teacher began to give Robbie special attention for studying (intervention phase), Robbie's study behavior increased dramatically. Withdrawal of the special attention (reversal phase) was accompanied by a reduction in

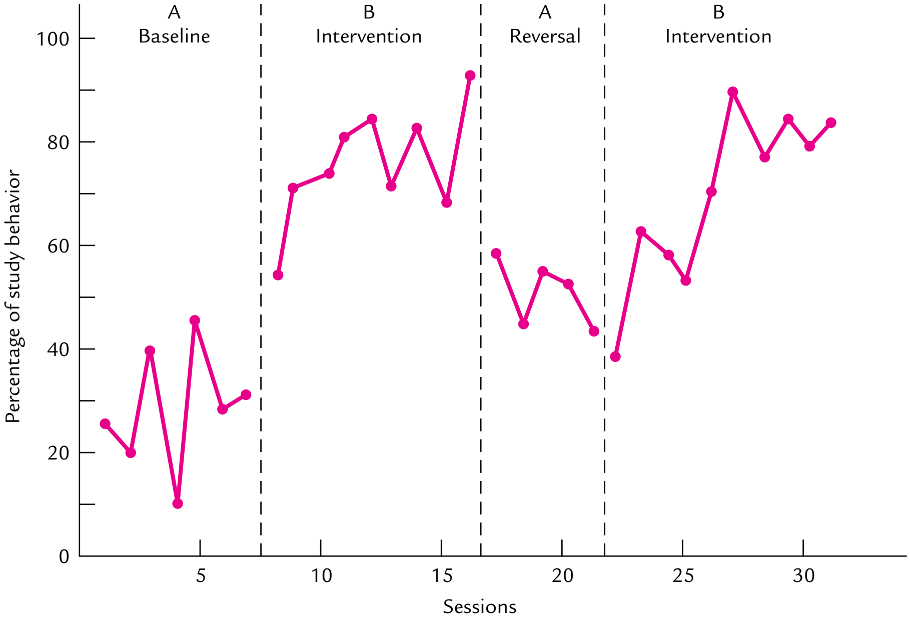


FIGURE 12-3 Percentage of study behavior during baseline and reinforcement showing intrasubject replication.
SOURCE: Hall, Lund, and Jackson, 1968; reprinted with permission.

studying, but the original baseline was not recovered. The return to the intervention phase brought a gradual return to previous reinforcement levels.

In this example, the baseline obtained during the original baseline phase was not recovered during replication. Nevertheless, the change in response rates from the baseline phase to the reinforcement phase is similar on both occasions, and there is little doubt that this change is reliable even if the amount of change is not.

Whether you would consider the intrasubject replication shown in Figure 12-3 successful would depend on your experimental question. If your question asked whether reinforcement increases the rate of studying, then the answer is yes, and the replication was successful. Studying increased relative to the baseline phase on both occasions. If your question asked by what amount studying increases, however, then the answer differs from first to second administration, and the replication was not successful.

Determining the Generality of Findings

As previously noted, single-subject baseline designs use intersubject replication (typically with three to six subjects) in order to establish whether the findings generalize across subjects. Intersubject replication does not always succeed.

A classic example of a failure of intersubject replication was provided by George Reynolds (1961). In this experiment, two pigeons were trained to peck at a translucent response key. Pecking on the key when it displayed a triangle against a red background led to occasional food reward. Pecking on the key when it displayed a circle against a green background led to nonreward. Both pigeons quickly learned to peck during triangle/red and not to peck during circle/green. They were then tested to see how much they would peck at the key when it displayed each stimulus shape or color separately. One pigeon pecked when the key displayed the triangle but not when it displayed the red color, even though both stimuli had been associated with reward during training. The other pigeon pecked when the key was red but not when the triangle was present.

To explain this failure to obtain intersubject replication, several theorists have suggested that each bird must have attended to different aspects of the original stimuli on which they had been trained. Apparently, one pigeon must have focused on the shapes and the other on the colors. The phenomenon has been termed *overshadowing* and has been used to support theories of selective attention.

Note that although the intersubject replication failed in the Reynolds (1961) experiment, the failure itself revealed a general principle: Learning to discriminate a complex stimulus on the basis of one aspect of it blocks learning about other, equally predictive aspects. The fact that the two birds' key-pecking behaviors came under the influence of different stimuli within the compound stimulus suggests that uncontrolled determining factors were at work. These factors could be the subject of further research. Such factors tend to be hidden by the averaging process when a group approach is used.

Intersubject replication establishes the generality of results across subjects, but establishing the generality of findings across experimental settings requires a different approach. Results are usually double-checked in new experiments. These experiments build on the original findings while extending the range of assessed variables, the types of variables manipulated, and/or the kinds of subjects tested. For example, the patterns of responding generated under various schedules of reinforcement have been replicated by using such varied reinforcers as food, water, chocolate milk, and cigarettes with such diverse subjects as goldfish, rats, pigeons, cats, dogs, monkeys, dolphins, and humans. Such extensions that incorporate aspects of the original experiment while adding new wrinkles are termed **systematic replications** to distinguish them from exact or **direct replications** (Sidman, 1960).

QUESTIONS TO PONDER

1. What factors affect your decision concerning choosing a stability criterion?
2. How is uncontrolled variability handled in the single-subject approach?
3. How do the single-subject and group approaches differ with respect to handling uncontrolled variability?
4. How is the generality of research findings established in single-subject research?

Dealing With Problem Baselines

In addition to excessive uncontrolled variability, problems you may have to deal with in baseline designs include drifting baselines, unrecoverable baselines, unequal baselines between subjects, and inappropriate baseline levels.

Drifting Baselines In some cases, it may prove impossible to stabilize a baseline against slow, systematic changes (drift). For example, during an experiment in which the dependent measure is basal skin conductance (a psychophysiological measure of arousal), conductance may gradually drift upward or downward as time passes during the experiment. If attempts fail to control this drift, you may be able to deal with the drift by effectively subtracting it out.

Figure 12-4 shows the results from a hypothetical ABAB experiment in which the baseline drifted systematically. Note that the baseline drifted gradually upward within each phase. Because the drift was consistent, it is possible to estimate the position (dashed lines) that the baseline would have reached at any point during the experiment had the treatment not been introduced. The effect of the treatment is clearly discernible after allowing for the drift.

Unrecoverable Baselines Another potentially serious problem arises if baseline levels of performance cannot be recovered during reversal. Such changes are considered carryover effects, the familiar problem discussed in Chapter 10 that plagues within-subjects designs. Some carryover effects render the baseline completely unrecoverable (in which case it becomes impossible to conduct a successful intrasubject replication). Special designs are required to deal with such completely irreversible changes. A discussion of these special designs appears later in the chapter. Other carryover effects are less of a problem because they render the baseline at least partially recoverable.

Such partially recoverable baselines frequently occur when learning develops during a treatment condition. In a simple operant conditioning experiment, for

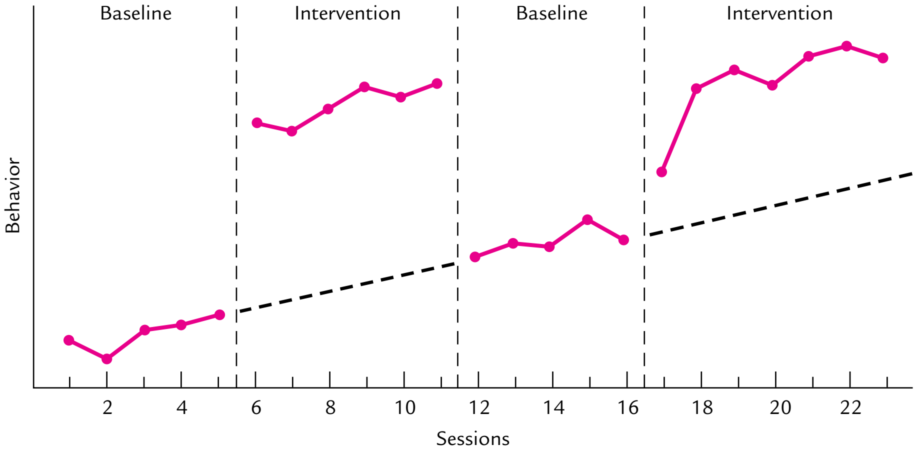


FIGURE 12-4 Hypothetical single-subject data showing a drifting baseline.

example, rats may rarely press a lever during the initial baseline phase (before reinforcement is introduced). During the reinforcement phase, subjects learn that pressing the lever produces food, and the reinforcing effect of this contingency generates high response rates. Return to baseline conditions at this point often fails to produce a return to baseline levels of responding.

Despite the lack of reinforcement for lever pressing, the rats repeatedly approach the lever and press it in a series of widely spaced "bursts" of responding that are characteristic of extinction conditions. As a result, the rate of responding (although considerably lower than that obtained during reinforcement) remains somewhat elevated relative to the initial baseline rate. The rats are no longer naive concerning the potential result of lever pressing.

Partial reversals such as this present few problems for analysis as long as a clear, replicable change remains in the levels of performance across treatments. You may even be able to remove some such carryover effects by taking appropriate steps. For example, if partial reversal results from fatigue or adaptation, you can minimize the effects of these variables by providing rest periods between your experimental treatments.

Unequal Baselines Between Subjects In some cases, the baselines of different subjects in an experiment level off at very different values even though the conditions imposed on the subjects are nominally identical. For example, after the same number of hours of deprivation, one rat may press a lever vigorously to earn a food reward whereas another may respond in a lackluster fashion. These initial differences in response rates may then produce different rates of learning in the treatment condition and result in apparently different functional relationships, a failure of intersubject replication.

In this case, identical levels of deprivation generate different levels of motivation because physiological differences exist between the subjects. To reduce the differences in motivation, you may increase the level of deprivation of the rat with the lower response rate. A little experimentation may provide a level that produces response rates similar to those of the first rat. With the baseline rates equated, the two subjects may now perform similarly across treatments.

Because the dependent variable is repeatedly measured during the baseline conditions, such steps may be taken to fine-tune the baseline to meet desired characteristics of stability and comparability. If comparable baselines are established across subjects, achieving intersubject replication may become more likely.

Inappropriate Baseline Levels Even if all subjects show similar baseline levels during the baseline phase, the particular levels obtained may not be useful for evaluating the effect of subsequent manipulations. A low baseline is desirable if you expect the treatment to increase the level of responding, but it is clearly undesirable if you expect the treatment to *decrease* the level of responding. Studies of the effects of punishment on behavior fall into the latter category. Detecting any suppressive effect of punishment would be difficult if the dependent variable were already near zero before the punishment contingency was introduced. Similarly, you will not be able to detect a facilitating effect of a treatment on behavior if the baseline starts at a value near its ceiling.

The solution to these problems is usually obtained by adjusting the experimental conditions to produce the desired baseline levels. In the punishment experiment, for example, you might increase the baseline response rates by reinforcing responses according to a variable interval schedule. You could then adjust the schedule to produce any desired level of responding. You would maintain the same schedule in the punishment condition. Thus, you could attribute any changes in the rate of responding to the punishment contingency.

QUESTIONS TO PONDER

1. What is a drifting baseline, and how can you deal with one?
2. What is an unrecoverable baseline, and what can you do if you have one?
3. What can you do if you have unequal baselines between subjects?
4. What can you do if you have an inappropriate baseline?

Types of Single-Subject Baseline Design

As yet, no widely accepted nomenclature exists to describe the wide variety of single-subject designs, although a few descriptive terms have emerged. This section differentiates among designs that manipulate a single independent variable (single-factor designs), those that manipulate two or more independent variables (multifactor designs), and those that measure several dependent variables (multiple-baseline designs).

Single-Factor Designs We have focused attention thus far on the ABAB design, which offers a complete intrasubject replication of the baseline (A) and intervention (B) phases of an experiment. Although less common, AB and ABA designs are also used. The AB design presents only a single administration of each condition and thus lacks intrasubject replication. Confounding by time-related factors is a serious problem with this design. The ABA design includes a reversal phase in which baseline conditions are reestablished after exposure to the treatment. This baseline reassessment allows you to determine whether the observed changes in behavior after treatment introduction were caused by the treatment. However, it lacks the final return to the B phase and thus fails to establish the recoverability of the baseline in the intervention phase. AB designs may be necessary if the intervention phase produces irreversible changes or (in an applied setting) if it is desirable to continue a treatment once it has been initiated. ABA designs may be appropriate if it is desirable to return the subject to preexperimental conditions prior to the termination of the study.

These basic procedures can be extended to include multiple levels of the independent variable. As in group designs, if these levels represent quantitative differences, the design is said to be *parametric*.

Using multiple levels of the independent variable presents certain problems for single-subject designs. Because only one or a few subjects are tested, completely counterbalancing the order of treatments across subjects is not usually possible. Instead, each subject may be exposed to the same order of treatments, but treatments will be presented repeatedly in different orders to assess the degree of carryover.

As an example of this counterbalancing strategy, consider a parametric single-factor experiment in which the three levels of the independent variable are A, B, and C. A single subject might be exposed to these treatments in the following order: A, B, A, C, B, C. This order provides transitions between close values of the independent variable (A–B, B–C, C–B), as well as a transition between distant values (A–C). Additional subjects might be tested with different orders. Note that this design provides a single replication of each treatment and thus represents a logical extension of the ABAB design.

Sometimes what starts out as an ABAB design may end up being extended and modified in the light of the initial findings. Kelly Therrien, David Wilder, Manuel Rodriguez, and Byron Wine (2005) ended up with an ABABC design in this way when conducting research in an applied setting. The study was carried out at one location of a sandwich restaurant chain to assess the effect of an intervention on the percentage of customers who were greeted by an employee within 3 seconds of entering the store. A greeting was defined as "any verbal acknowledgment given by an employee" (e.g., "Hello" and "What can I get you?). Three or four employees were behind the service counter at all times.

The intervention consisted of having the manager behind the counter and a bell that rang each time the door opened. Data were collected by observers, posing as customers, who sat at a nearby table and surreptitiously recorded the percentage of customers who were greeted by an employee (other than the manager) within 3 seconds of entering the shop. Data were first collected during a baseline (A) phase in which the intervention was absent. After seven daily baseline sessions, the "manager + chime" intervention (B) phase was imposed. This continued for five sessions and was followed by a return to the baseline (A) phase. After a further six sessions in the baseline phase, the intervention (B) phase was imposed again. Because the level of performance during the intervention was not as high as desired, a final (C) phase was added that continued the intervention but added feedback to it. During this phase, immediately following each session, the manager graphed each employee's greeting performance and individually showed each employee the graph. The employee was praised if the new point was higher than the previous one (e.g., "You guys look great!") but said nothing otherwise.

Figure 12-5 shows the results averaged across employees. (Results were said to be similar for most individuals.) The data clearly show that the intervention was effective in raising the level of customer greeting and that this effect was reversible. The addition of feedback seems to have brought about additional improvement, although the reliability of this effect was not assessed in the study through replication.

This applied study did not employ a stability criterion to determine when to move to the next phase. (As noted previously, often in real-world settings, factors such as time constraints limit the ability to impose such a criterion.) Even so, the effects of the interventions are clearly evident in the graphs.

If you must test a number of levels of your independent variable and are concerned about possible drift in your baseline, you may want to include a return to the baseline phase after each exposure to a treatment phase. One of us (Abbott) and

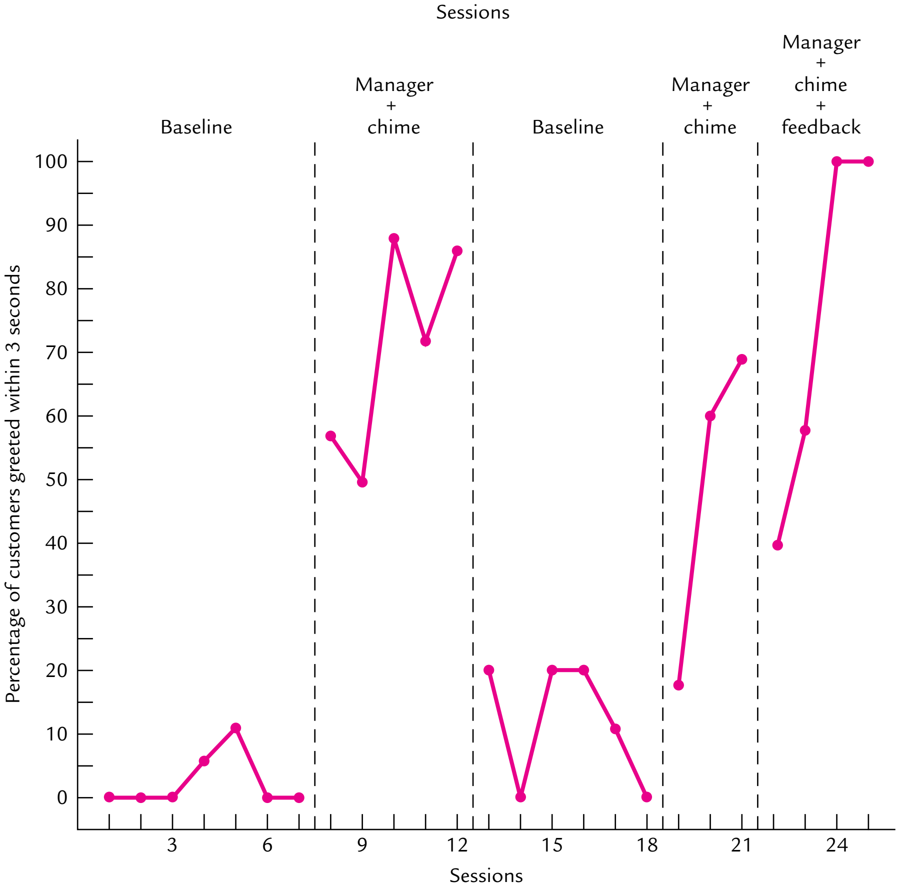


FIGURE 12-5 Results of the customer-greeting study for the baseline and intervention phases. The original ABAB design was extended by the addition of a final manager + chime + feedback phase.

SOURCE: Therrien, Wilder, Rodriguez, and Wine, 2005; reprinted with permission.

Pietro Badia (1979) used this technique with rat subjects. The experiment assessed the preference for signaled over unsignaled foot shocks. Signaled shocks were preceded by a warning tone whereas unsignaled shocks were not. Subjects could choose to receive signaled shocks by pressing a lever. Each response produced 1 minute in the signaled shock schedule, identified by illumination of a house light. At other times, unsignaled shocks were delivered.

The independent variable was the length of the signal, which was systematically varied across treatments from 0.5 second to 2.0 seconds (in half-second steps). Baseline response levels were collected during training phases, which also familiarized subjects with the signaled and unsignaled schedules prior to each test phase. Each training phase provided a given signal length and was followed by a test phase at the same length.

Figure 12-6 shows the results of the Abbott and Badia (1979) experiment. The dependent variable (percentage of time in changeover) reflects the number of minutes in the signaled condition "bought" by responses as a percentage of total session time. During training, the responses actually had no effect, but the number of minutes that would have been bought by these responses served to provide the baseline. The average percentage of time in changeover is shown across the final three stable sessions in each condition during the baseline (the black line) and testing (the red line) phases as a function of signal length. Each graph represents the performance of a different subject.

The levels of changeover responding during testing were low (with the exception of one subject) during initial testing at a signal length of 0.5 second. These levels increased as the signals were lengthened across treatments until they reached nearly maximum values. When signals were then shortened, responding tended to decline. These changes were not caused simply by baseline drift. This is shown by the fact that the baselines collected during training phases remained relatively low throughout the experiment.

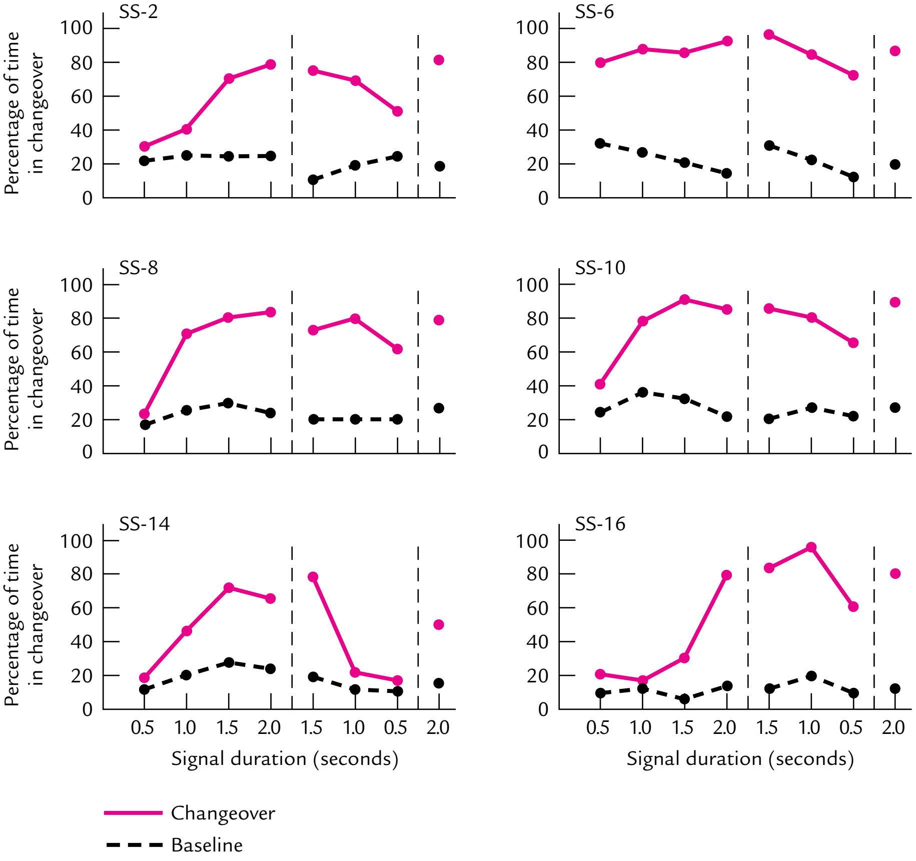


FIGURE 12-6 Results of an experiment in which baselines (the dashed lines) were repeatedly assessed. Each graph represents the performance of a single subject.
SOURCE: Abbott and Badia, 1979, p. 413; reprinted with permission.

Also note the failure to recover the original levels of responding on the final return to the 0.5-second signal length in four subjects (all but SS-6 and SS-14). This indicates that some carryover effects were present. Even so, the effect of signal length on changeover responding is clear. Shorter signals supported weaker responding than did longer signals.

Multifactor Designs Single-subject designs can include more than one independent variable. As in factorial group designs, you can assess the main effects of independent variables and their interactions. Murray Sidman (1953) provided an example of this design during an investigation of responding on the free operant avoidance schedule that bears his name. The Sidman avoidance schedule delivers a brief shock at regular intervals (e.g., every 5 seconds) during the shock–shock (S–S) interval. A response, usually a lever press, terminates the S–S interval and starts another interval, the response–shock (R–S) interval. If the subject fails to respond during the R–S interval, a shock occurs and a new S–S interval begins. If the subject responds during the R–S interval, no shock is delivered, and the subject is returned to the beginning of the R–S interval. Thus, if the subject always responds during the R–S interval, all scheduled shocks can be avoided.

Sidman (1953) investigated the effect of varying the length of both the S–S and R–S intervals. Thus, this experiment had two factors: length of the S–S interval and length of the R–S interval. Rats were exposed to several levels of each independent variable in every combination.

Figure 12-7 shows the results from this two-factor single-subject experiment. The figure shows the rates of lever pressing generated under each combination of S–S and R–S intervals. The graph shows portions of each condition in which the stability criterion was met. As you can see, the number of responses per minute was affected by the lengths of both intervals.

Because the data from multifactor single-subject experiments are not submitted to a statistical analysis, omitting some cells of the factorial matrix presents no special analytical problems. If the functional relationships between independent and dependent variables follow regular patterns (as is usually the case), then it is possible to "sweep out" the functions. You can do this by providing data points at well-placed intervals rather than at every possible combination of levels. Each subject must be exposed to every combination of levels for which a point is required. Using less than the full factorial number of combinations can result in considerable savings in the time required to complete the experiment.

Multiple-Baseline Designs Some treatments cause irreversible changes in behavior, and a special approach is required to deal with this problem. **Multiple-baseline designs** provide one solution. These designs simultaneously sample several behaviors within the experimental context to provide multiple behavioral baselines.

As an example, imagine you have developed a new technique for eliminating undesirable habits. You expect the changes that it produces to be relatively permanent, so a reversal design is clearly inappropriate. You decide to use a multiple-baseline design.

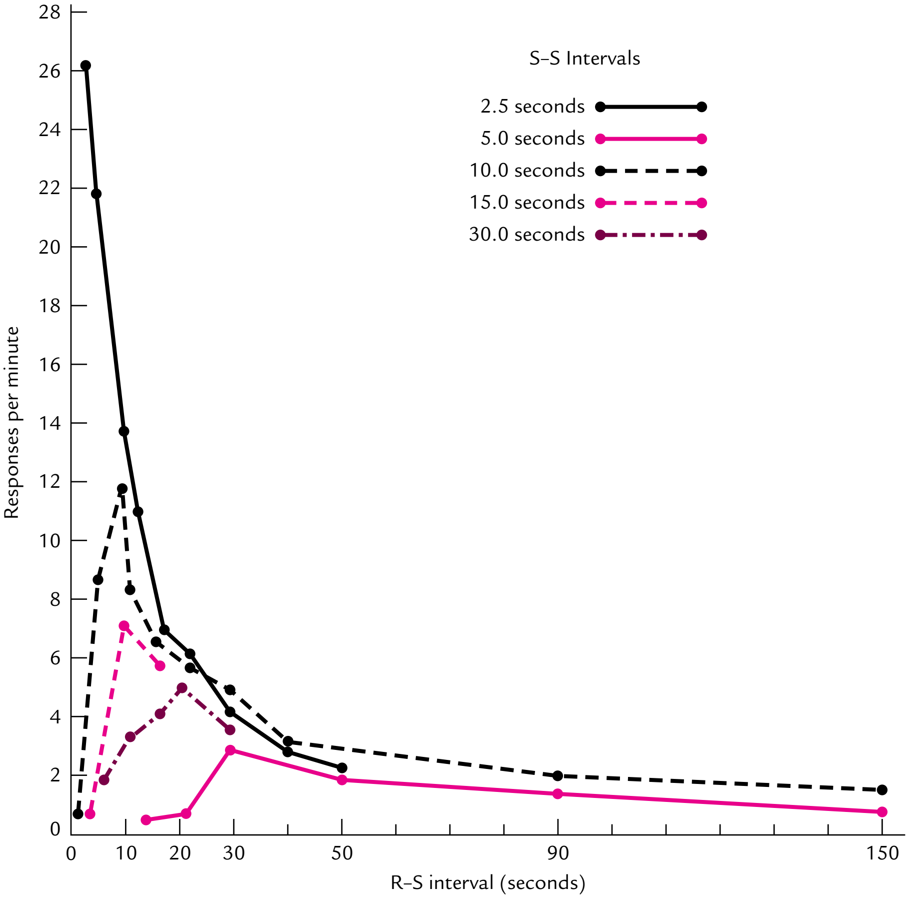


FIGURE 12-7 Performance of a single subject in a two-factor design in which both R–S and S–S intervals were manipulated.
SOURCE: Sidman, 1953; reprinted with permission.

To test your technique, you identify a few individuals who have at least two undesirable habits they wish to kick: smoking and excessive coffee drinking. You begin your study by simply observing and recording the frequencies of these two behaviors across a number of days to establish a baseline for each behavior. You then introduce your treatment but apply it to only *one* of the behaviors. For one subject, you choose to attack smoking and for the other, coffee drinking.

The treatment appears to be successful. The treated behavior soon declines to levels well below the baseline. At the same time, however, the untreated behavior remains at its previous baseline levels. When the new levels of the treated behaviors stabilize, you begin to apply the treatment to the remaining behaviors. These, too, now decline to low levels.

Figure 12-8 shows the data from the hypothetical multiple-baseline study. Note that each behavior changes only after the treatment is introduced for that behavior. Which behavior is treated first apparently makes little difference.

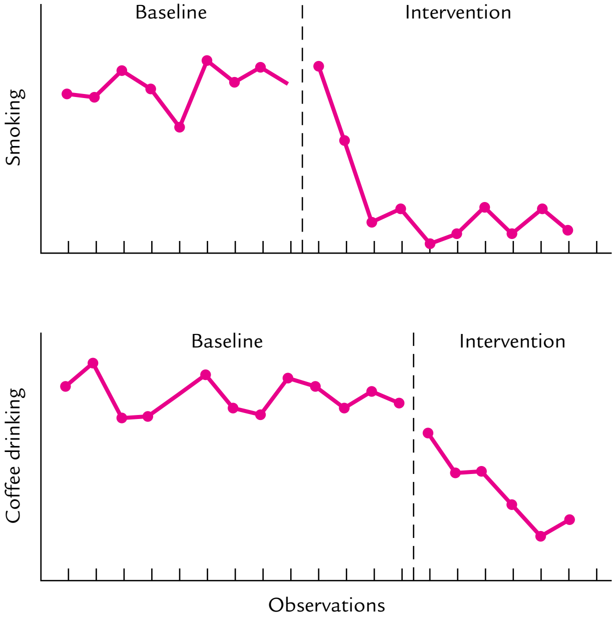


FIGURE 12-8 Results of the hypothetical multiple-baseline study. Both behaviors are collected simultaneously from a single subject.

The multiple-baseline design uses the untreated behavior as a partial control for time-correlated changes that may confound the effect of the independent variable. It is possible that the change in the treated behavior would have happened when it did even if the treatment had not been introduced. However, if this change was not caused by the treatment, the untreated behavior likely would have changed as well. In addition, the untreated behavior most likely would not subsequently change as soon as the treatment was applied to it.

For the multiple-baseline design to be effective, the behaviors chosen for observation should be relatively independent of one another. If the behaviors are correlated, then applying the treatment to one of the behaviors will affect both. Your ability to discriminate treatment-induced changes from changes induced by time-correlated confounding factors will be seriously hampered.

Debbie Westerlund, Elizabeth Granucci, Peter Gamache, and Hewitt Clark (2006) used the multiple-baseline design to assess the effectiveness of peer mentoring of four female cosmetology students afflicted with specific learning disabilities and/or severe emotional disturbances. For each student, the researchers identified two behaviors in need of intervention. Two students received peer mentoring for hair-roller setting (the proper steps followed, rollers correctly sized for the hairstyle, and hair properly distributed among the rollers, etc.) and for combing out (the proper steps followed, hair combed out properly for the hairstyle, etc.). The other two students received peer mentoring for comfort inquiry (asking the client such questions as “Is the water temperature comfortable?”) and for suggestion statements about such things as hairstyling products and recommended haircuts.

The peer mentors assisted the students in ways designed to minimize embarrassing them. Such assistance included demonstrations, corrective feedback offered as suggestions, and verbal prompts delivered quietly or away from patrons. The mentors explained or modeled the behavior, allowed the student to demonstrate the behavior, and provided descriptive praise for portions of the behavior that were correctly demonstrated.

The multiple-baseline design was implemented in three phases. In the first, baseline phase, target behaviors were monitored, and their levels recorded over several sessions. In the second phase, the intervention commenced for one of the two target behaviors. Several sessions later, intervention commenced for the second target behavior.

Figure 12-9 shows the results for one of the students. (Similar results were obtained for the other three.) The graph shows the percentage of steps completed for roller setting and for combing out during the baseline and intervention sessions. As the figure indicates, the percentage of steps completed was low for both behaviors during the baseline phase. The intervention for roller setting commenced on Session 5 and produced an immediate and substantial improvement in performance that continued over the remainder of the sessions. Meanwhile, the second behavior, combing out, continued to be performed poorly. On Session 7, the intervention for

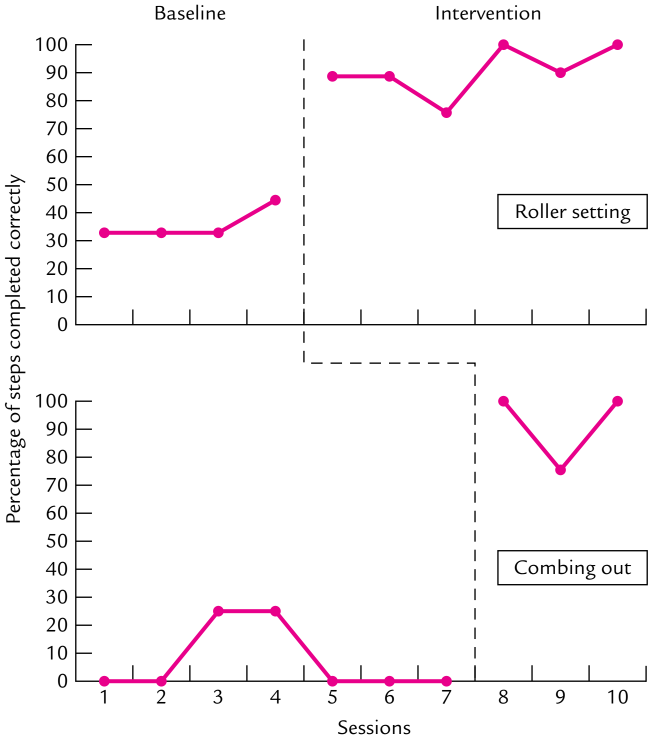


FIGURE 12-9 Results of the peer-mentoring multiple-baseline study for one participant.
SOURCE: Westerlund, Granucci, Gamache, and Clark, 2006; reprinted with permission.

combing out began, and this behavior too showed an immediate and substantial benefit. In each case, the improvement in a behavior began only after peer mentoring had commenced for that behavior, so it is unlikely that the timing of these changes is coincidental.

QUESTIONS TO PONDER

1. What are the characteristics of the single-factor baseline design?
2. What are the characteristics of the multifactor baseline design?
3. What is a multiple-baseline design, and when would you use one?

DYNAMIC DESIGNS

Although baseline designs afford the opportunity to examine moment-to-moment changes in behavior within each baseline or treatment phase, their primary use is to establish how behavior differs from one level of an independent variable to another in the steady state. Adaptation to new conditions may require time and experience; if so, behavior observed immediately after a switch from one treatment to another may not typify the stable pattern that may emerge after more extensive exposure to the new treatment. For this reason, subjects are kept under each treatment condition until behavior shows no sign of further systematic change.

This emphasis on steady-state behavior fosters the use of designs in which independent variables are manipulated in discrete levels even when the variable itself is continuous. For example, the key-pecking behavior of pigeons may be examined at several widely separated levels of food deprivation (e.g., 80%, 90%, and 100% of free-feeding body weight). Subjects are maintained at each level of deprivation until their behavior stabilizes and then are moved to the next level. In such designs, behavior immediately following the change in level (described as *transitional* to distinguish it from steady-state behavior) can be observed to determine what has been called *behavioral dynamics*—regular patterns of behavioral change over time. A nice example of this approach is provided by William Palya, Don Walter, Robert Kessel, and Robert Lucke (1996), who investigated behavioral dynamics following an unsignaled step transition from variable-interval reinforcement to extinction in pigeons. To emphasize the regularities in these dynamics, the curves relating response rate to time following the transition were averaged across repeated transitions (but not across subjects). Figure 12-10 depicts the reinforcement rates and response rates as functions of time (in seconds) for four birds. The step transition to extinction occurs at 200 seconds in each graph. Each bird showed a rapid decrease in response rate, which began within seconds of the transition.

Step changes in the level of the independent variable may reveal interesting regularities in transitional behavior, but it may be at least as informative to record behavior during *continuous variation* of the independent variable as long as (1) the rate of variation is not so fast that the behavioral changes cannot "keep up" and (2) the changes are more or less reversible. Such a design lacks the discrete values

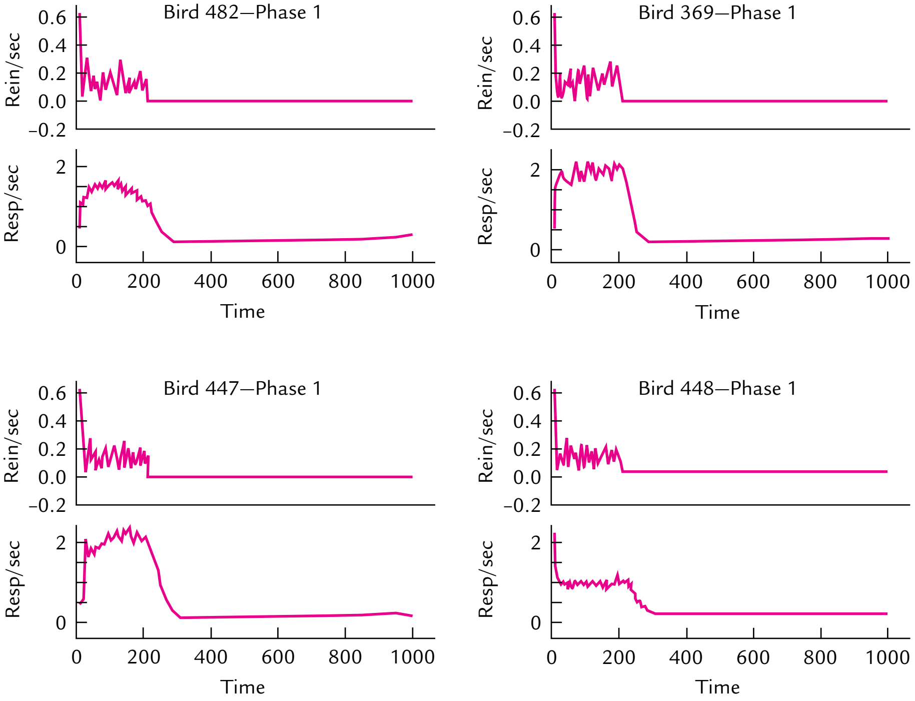


FIGURE 12-10 Step-transition data from four pigeons.
SOURCE: Palya, Walter, Kessel, and Lucke, 1996; reprinted with permission.

of the independent variable that serve to distinguish the baseline and intervention phases of the baseline design, so, strictly speaking, it may not be appropriate to refer to studies using continuous independent variable variation as baseline designs. For this reason, we have chosen to identify designs that include a continuously varying independent variable as **dynamic designs**.

A typical dynamic design was used in a "compensatory tracking" experiment described by Powers (1978). An individual participant was given the task of keeping a cursor (a short, vertical line) aligned with a target (another short, vertical line), which were simultaneously presented on a computer screen. Although the target remained fixed in position on the screen, the cursor could be moved left and right by manipulating a joystick. However, an invisible force or "disturbance" seemed to be acting on the cursor, causing it to drift erratically left and right on its own. To keep the cursor on the target, the participant had to compensate for the cursor's drift by moving the joystick.

The independent variable in this experiment was the disturbance, which varied smoothly and continuously in size and direction over a programmed range of values. The continuously monitored dependent variable was the position of the joystick. Together with the target position, these values were used to compute the moment-by-moment position of the cursor and the error (difference between target and cursor positions).

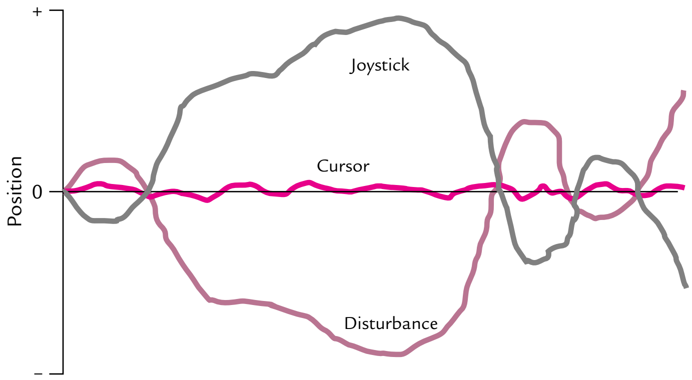


FIGURE 12-11 Continuous data from a dynamic design. The participant moved a joystick to keep a cursor horizontally aligned with a target mark at position zero over a period of 60 seconds (see text). Horizontal scale indicates time; vertical scale indicates vertical distances from target mark.

SOURCE: Adapted from Powers, 1978; reprinted with permission.

Figure 12-11 shows the data from one 60-second experimental run. One line shows the variation in the disturbance (independent variable) over the course of the run. The second line shows the position of the joystick, scaled to screen coordinates. Notice how this line is almost a mirror image of the first. As the cursor moved in one direction, the participant had to move the joystick in the opposite direction in order to cancel out the cursor's movement and keep it over the target. The third line (which varies over a much smaller range than the other two) shows the position of the cursor relative to the target. The fact that cursor excursions were generally small shows that the participant was highly successful in keeping the cursor near the target. The data from such experimental runs were used to evaluate a mathematical model (derived from control theory) that used the moment-by-moment disturbance values to predict the participant's joystick movements. The joystick positions predicted by the model matched the observed positions almost perfectly.

Although we have distinguished between baseline and dynamic designs, it is important to understand the close relation between the two types. The Palya et al. (1996) design described earlier can be thought of as a baseline design involving discrete values of the independent variable. However, it also could be viewed as a dynamic design in which the independent variable changes as a step function, swinging instantly between extreme values. More important than the label you choose to describe the design is the kind of information the study was designed to collect. In the Palya et al. study, the focus was on the dynamics of behavioral change.

The method of continuously varying the independent variable while observing continuous, dynamic changes in the dependent variable has been used infrequently in the behavioral sciences but should become more common as interest in behavioral dynamics increases. In May 1992, the *Journal of the Experimental Analysis of Behavior* devoted an entire issue to the topic of behavioral dynamics. There you can find a number of examples of this continuous approach to single-subject research.

DISCRETE TRIALS DESIGNS

Although baseline and dynamic designs can be powerful tools for discovering causal relationships in the single subject, they will not work in every experimental situation. Imagine, for example, that you are interested in studying the ability of an air traffic controller to detect a radar signal representing a single airplane in trouble (the "signal") against a radar screen full of radar signals from other airplanes (the "noise"). In one condition of your experiment, you present the target radar "blip" embedded within other radar blips ("signal + noise," in signal-detection theory terms). In the second condition, the target blip is absent (only the other radar blips are present—"noise only"). After each trial, your participant indicates whether the target blip was present on the radar screen.

Using the baseline approach, you should first expose all your participants to a series of signal + noise trials. After every 50 trials, you calculate the number of yes responses and plot these numbers to provide the behavioral baseline. You then continue the signal + noise treatment until the baseline stabilizes and then switch over to the noise-only treatment. Can you see any problem with this design?

If you said yes, you're right! During the signal + noise treatment, your participant would soon begin to suspect that the target radar blip (the signal) was present on every trial. Before long you would discover that your baseline had jumped to 100% yes responses and would stay there, inducing a ceiling effect. The baseline would not provide a true reflection of your participant's ability to detect the target blip because on many of the trials the participant would respond yes simply out of habit and not because the signal had actually been detected.

In such cases, the baseline approach must be abandoned in favor of a design that will discourage the participant from establishing a response set, as was the case in the previous example. Fortunately, such a design exists: the **discrete trials design**. Like baseline designs, discrete trials designs focus on the behavior of the individual participant (e.g., your air traffic controller) rather than on group behavior.

Characteristics of the Discrete Trials Design

Both the baseline and discrete trials designs seek to rigidly control extraneous sources of variance, to make the effect of the independent variable readily visible. Unlike the baseline design, however, the discrete trials design does not produce a continuous within-treatment baseline that can be adjusted and fine-tuned. Instead, behavior measured over a series of discrete trials must be averaged to provide relatively stable indices of behavior under the various treatment conditions. The major characteristics of the discrete trials design are shown in Table 12-2.

The single-subject designs commonly used in experimental psychology prior to the 1920s were generally discrete trials. They continue to be used today, especially in psychophysics (which studies the relationship between physical stimuli and the sensations they generate), as well as in some areas of human judgment and decision making.

An example of such a design is provided by an experiment on signal detection reported by Wilson Tanner, John Swets, and David Green (1956). The problem in

TABLE 12-2 Characteristics of the Discrete Trials Design

1. Individual subjects receive each treatment condition of the experiment dozens (perhaps hundreds) of times. Each exposure to a treatment, or trial, produces one data point for each dependent variable measured.
2. Extraneous variables that might introduce unwanted variability in the dependent variable are tightly controlled.
3. If feasible, the order of presenting the treatments is randomized or counterbalanced to control order effects.
4. The behavior of individual subjects undergoing the same treatment may be compared to provide intersubject replication.

signal detection is to determine how good a given observer is at detecting a signal that may be almost buried in noise. This is traditionally indexed by the observer's "hit rate," which is the proportion of trials on which the signal was present and the observer reported detecting the signal. However, the hit rate is affected by more than the observer's ability to separate the signal from the noise. It also is determined by the observer's *response bias*, or willingness to decide that a signal was present when he or she is uncertain about that. Observers with a "liberal" response bias will tend to guess yes whereas those with a "conservative" response bias will tend to guess no. Guessing yes increases the hit rate, so, everything else being equal, those with a liberal response bias will *seem* to be better at detecting the signal than those with a conservative response bias. Thus, in traditional experiments, the ability to detect the signal is confounded by response bias.

The Tanner et al. (1956) experiment was designed to eliminate this confound by separately measuring the observer's response bias and ability to detect the signal. Part of the strategy involved measuring not only the hit rate but also the "false-alarm" rate. This is the proportion of trials on which the signal was *not* present and the observer decided that the signal *was* present. Guessing yes when uncertain not only increases the hit rate (on trials when the signal was present) but also increases the false-alarm rate (on trials when the signal was not present).

In the experiment, two participants were exposed to a series of trials in which an auditory signal was either present or absent against a background of noise. On each trial, the participants were required to respond yes if they thought they heard the signal and no if they did not.

The participants' response biases were manipulated by systematically varying the probability of the signal being present on a trial. When the probability was low, participants would most often be right when guessing if they guessed no and therefore adopted a conservative response bias. When the probability was high, participants would most often be right when guessing if they guessed yes and therefore adopted a liberal response bias. In the experiment, the probability of signal presentation was systematically varied across days from 0.1 (1 out of 10 trials) to 0.9 (9 out of 10 trials). Participants received 300 trials per day and spent 2 days at each probability level.

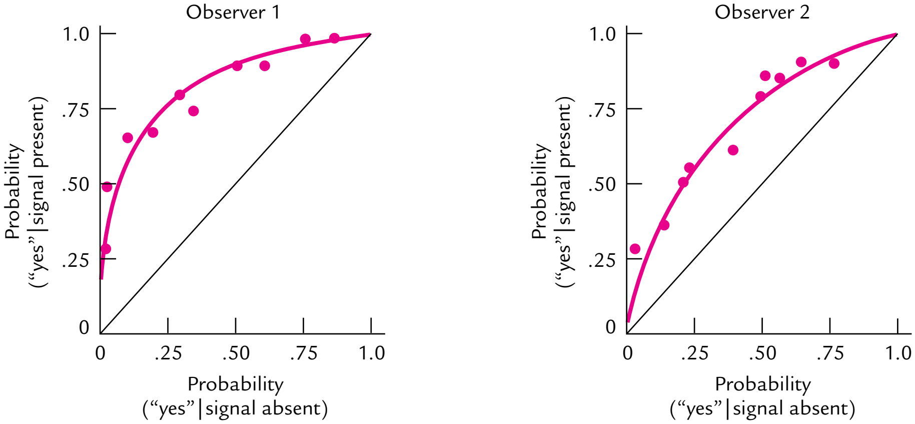


FIGURE 12-12 Results for two observers in a signal-detection study.
SOURCE: Based on data from Tanner, Swets, and Green, 1956.

Figure 12-12 shows the results for each participant. The figure depicts the hit rate (probability of saying yes when the signal was present) plotted against the false-alarm rate (probability of saying yes when the signal was absent). The diagonal line represents the points at which hits equal false alarms, a situation that indicates no ability to detect the signal. Points falling above the diagonal indicate cases in which the hit rate exceeded the false-alarm rate and thus demonstrate some ability to detect the signal above the noise. All points falling along the same curve indicate the same sensitivity to the signal, and curves lying farther above the diagonal indicate greater sensitivity than those lying closer to the diagonal. The position of a point along a given curve indicates the response bias, ranging from conservative (closer to the lower left) to liberal (closer to the upper right).

Analysis of Data from Discrete Trials Designs

Analysis of data from discrete trials single-subject experiments usually begins by averaging the responses across the repeated presentations of a particular treatment. A large number of presentations helps ensure that the resulting mean provides a stable and representative estimate of the population mean (i.e., of the mean that would be obtained if an infinite number of trials could be given to the same subject under the treatment conditions). The means obtained from the different treatment conditions may then be compared to determine whether they appear to differ. This comparison may or may not include assessment through inferential statistics to determine whether the observed differences are reliable.

The analysis applied to the data of discrete trials single-subject experiments is usually determined by a theory or model of the behavior being examined. For example, in the area of human judgment and decision making, a lens model analysis has often been applied to data collected from single subjects. Another example is provided by the theory of signal detectability that provided the analytical model for the signal-detection experiment previously described. Often these analyses yield a small number of descriptive statistics, such as the *d*′ (a measure of sensitivity) and ß (a measure of response bias) in signal detection. If a relatively large number of subjects have been tested, these descriptive measures (although derived from a single-subject analysis) may be used as data for a between-subjects analysis.

Used in this way, inferential statistics are not applied to analyze the data of an individual subject but groups of subjects. Some investigators have argued for the application of inferential statistics to single-subject data (see Kazdin, 1976).

INFERENTIAL STATISTICS AND SINGLE-SUBJECT DESIGNS

Those who advocate the application of inferential statistics to data from single-subject designs would not want to use them as a substitute for control over variables and replication. However, in some cases, they argue that the desired level of control is difficult to achieve (e.g., in some clinical situations). For these cases in which the necessary control cannot be obtained, they suggest that inferential statistics may provide a solution.

If you choose to take this route, be aware that the usual statistical procedures developed for group designs cannot be applied to data from single-subject designs without modification, and even then problems may exist (Kazdin, 1976).

The most straightforward approach is to use the multiple observations taken within a treatment to provide an estimate of uncontrolled error variance (in group designs, this estimate is provided by within-treatment observations of multiple subjects). This estimate is then compared with the variance of scores between treatments to give an estimate of the probability that the treatment deviations were the product of chance (Chassan, 1967).

Unfortunately, this approach is upset by the serial dependency across data points. That is, scores from adjacent observations within a treatment are more likely to be similar to each other than to scores from more widely separated observations. This is similar to the problem of correlated scores that appears in within-subjects designs (see Chapter 10). For reasons not discussed here, it is a more serious problem for single-subject designs. Several ways to deal with serial dependency have been proposed, but none is entirely satisfactory. In fact, Manolov, Solanas, Bulté, & Onghena (2010) state that "[h]ow to analyze single-subject data is a question yet to be answered." These authors found in a study of simulated data from ABAB designs that randomization test results were distorted by the presence of autocorrelation (a measure of serial dependency) and strongly affected by the relative lengths of the four phases. For a discussion of the general problem and a review and assessment of options available at the time, see Kazdin (1976, 1978). See Manolov et al. (2010) for a more contemporary discussion of these issues.

ADVANTAGES AND DISADVANTAGES OF THE SINGLE-SUBJECT APPROACH

A design that affords advantages in one arena often carries disadvantages in another. In this section, we explore some of the major advantages and disadvantages of the single-subject approach.

The main advantage of the single-subject approach is its focus on controlling error variance. By focusing on behavior of individual subjects (as opposed to looking at group means), you may better identify potential sources of error and control them. On a related note, focusing on individual subjects may lead to a truer estimation of the impact of the independent variable. Individual patterns of behavior often reveal nuances obscured by averaging used in the group approach. (This issue is discussed again in more detail in Chapter 13.)

Single-subject designs in general (and baseline designs in particular) require that a single subject's behavior be followed over a relatively large number of observations. A perhaps extreme example was provided by an experiment conducted by one of the authors of this text (Abbott) in which the same rat subjects were tested in various conditions of the experiment for more than a year. This experiment included two independent variables. Each assessment phase was preceded by a training phase in which baselines were established. About nine days, on the average, were required to reach the stability criterion for the behavior, and each phase had to be replicated.

Of course, not many single-subject designs require this much time to test a subject, but most do require much more time than the equivalent group design would. This intensive investigation of the behavior of a single subject is both a strength and a weakness of the single-subject approach. It is a strength in that the long observational period often reveals nuances of behavior that might be missed in a short-term design. The ability to adjust and fine-tune the baseline over time provides an extended opportunity to identify previously unsuspected important variables. The intensive investigation is a weakness in that the strategy commits the investigator to a relatively long-term project that may be disrupted by uncontrolled factors such as illness or breakdowns of equipment.

On the bright side, even the breakdown of equipment can sometimes lead to new discoveries. Sidman (1960) documents how a sticking relay in his operant conditioning equipment resulted in the accidental delivery of unavoidable shocks to a rat that had been successfully responding to an avoidance schedule. The baseline of responding had been stable, but now began to climb to high rates. The finding was unexpected and generated a whole new direction of research.

Another advantage of single-subject designs is the fact that causal relationships can be established by using even just one subject. This is particularly important in clinical settings in which the objective of the research may be to identify an effective treatment for the behavioral disorder of a specific client. When more than one subject is used, the single-subject design permits the investigator to compare individual responses to the independent variable. Each subject may be found to exhibit reliable yet idiosyncratic responses to the same variable. Comparing subject-related variables may then suggest which differences between the subjects might be responsible for the differential responses. A design with only one subject may provide a good demonstration that a variable has an effect on behavior. However, to elevate the demonstration to the status of a general finding, the results should be replicated with additional subjects.

A disadvantage of single-subject designs is that the design is inappropriate for many research applications. For example, potential carryover effects may confound the effects of the independent variable. If carryover effects are severe enough to cause irreversible changes, then the single-subject approach may have to be abandoned in favor of the between-subjects approach. In addition, some research questions simply do not lend themselves to a single-subject design. Much of the research in social and developmental psychology could not be run as single-subject designs.

Moreover, the results from single-subject designs sometimes are of limited generality. By tightly controlling the experimental situation to reduce error variance, you may be creating a highly artificial environment within which the behavior of interest is observed. For example, if you study the effects of positive reinforcement on the behavior of autistic children in a tightly controlled laboratory, you cannot be sure that your results will apply to autistic children in a hospital setting. In short, tight experimental control of extraneous variables increases internal validity. However, remember that when you increase internal validity, you often reduce external validity.

A final disadvantage of the single-subject approach is that, despite all attempts to control extraneous variables, some variables cannot be easily controlled. Subject variables such as personality and intelligence cannot be controlled by tight experimental design. In research in which a homogeneous strain of rats serves as the subject population, this problem may not be important. However, when you apply single-subject research design to humans (such as in a clinical setting), these variables may come into play, and there is no easy way to eliminate their effects. The only option may be to measure those variables and statistically control their effects.

Consider single-subject research as an alternative to group-based research when appropriate. You should not try to fit every research question into a single-subject design. However, keep this option open when your research question could be best answered with a single-subject approach. Finally, even if you use a group design, you can apply some of the logic of the single-subject approach. Do not ignore individual subject behavior and look only at average performance. In many cases, looking at the scores of individual subjects within your experiment can help you to interpret your data.

QUESTIONS TO PONDER

1. What are the characteristics of a dynamic design?
2. What are the major characteristics of the discrete trials design?
3. How are inferential statistics used in single-subject designs?
4. What are the advantages and disadvantages of the single-subject approach?

SUMMARY

Single-subject (or small-*n*) designs allow you to establish causal relationships among independent and dependent variables while focusing on the behavior of one or a few subjects. These designs work by collecting large numbers of observations of a subject's behavior within each treatment condition, rigidly controlling extraneous variables that might contribute to unwanted variability in the dependent measure, focusing primarily on relatively powerful independent variables whose effects (given the small amount of uncontrolled variation in the dependent measure) are easily detected by inspecting graphs, and repeating each treatment condition to establish reliability through intrasubject replication.

Single-subject designs dominated psychological research prior to the 1920s but were overshadowed by group-based designs following the development of inferential statistics. Experiments based on single-subject designs were difficult to publish in the past, but today the single-subject approach has regained wide acceptance. This acceptance has arisen especially in applied areas such as clinical psychology in which the primary interest is in assessing the effectiveness of therapeutic procedures on individual clients.

Single-subject designs fall into three broad categories: baseline designs, dynamic designs, and discrete trials designs. Baseline designs repeatedly record the subject's score on the dependent measure within each treatment exposure to plot a baseline for the behavior. Steps are taken to identify and control extraneous variables so that the variability and drift of the baseline are reduced to the minimum level. Usually the baseline must meet a stability criterion before the next treatment condition can be introduced. Each treatment is repeated at least once (intrasubject replication) to determine the reliability of the findings. If more than one subject is tested (intersubject replication), comparison across subjects is performed to establish the generality of results. Generality and reliability are also assessed through systematic replication. This replication examines the same independent and dependent variables under conditions somewhat different from those of the original experiment (for example, different species of subject or different reinforcer).

Baseline designs include single-factor designs (which include one independent variable), multifactor designs (two or more independent variables), and multiple-baseline designs (more than one dependent variable). Single-factor designs may be of the AB, ABA, or ABAB type. The AB type evaluates behavior during a baseline (A) condition and then in a treatment (B) condition. The lack of intrasubject replication makes this design subject to confounding by time-related factors and therefore undesirable. The ABA design controls time-related factors by adding a second baseline evaluation after the treatment evaluation. The ABAB design provides a complete intrasubject replication of the experiment. The ABAB design is preferable to the ABA design, especially if it would be unethical to end the experiment after returning the subject's behavior to an undesirable state in the second baseline evaluation. The ABAB design format can be extended for multilevel variables. When such variables are quantitative, the design is said to be parametric.

Multifactor baseline designs require that different combinations of the independent variables be tested across the study. A factorial design may be used (in which every combination is evaluated) or specific combinations of interest may be tested. In either case, each treatment is evaluated at least twice to provide intrasubject replication. These designs can become extremely time consuming to conduct if the number of variable combinations to be tested is large.

Multiple-baseline designs provide a partial solution to the problem of irreversible treatment effects. Different behaviors are observed, and a baseline established for each. The treatment is then introduced separately for each behavior in staggered fashion across time. The treatment is judged effective if the level of each behavior changes only after the treatment is applied to it. The multiple-baseline approach requires that each behavior be relatively independent of the others.

Dynamic designs are similar to baseline designs but employ a continuously varying independent variable. The behavior of interest is monitored continuously to determine the dynamic response of the behavioral system to the ongoing changes in the independent variable. Such designs focus more on transitions of behavior as opposed to the steady state.

Discrete trials designs expose subjects to a series of trials, with each trial providing one exposure to a given treatment and yielding one score for each dependent measure. Each treatment is repeated a large number of times to provide stability to the treatment means. These designs are used in areas such as psychophysics, human judgment, and decision making, in which the interest focuses on the perceptual or decision-making abilities of individuals. Such designs often follow from a theoretical analysis of the behavior and yield descriptive statistics that are then evaluated in light of the theory. If a relatively large number of subjects are thus tested, the summary statistics from each subject may provide the data for a group-based inferential statistical analysis.

Some interest has recently been expressed in developing inferential statistics that can be applied to single-subject data. These statistics would be employed when the data cannot be stabilized or the independent variables are too weak to produce results that can be visually analyzed. A problem with such analyses is the serial dependency (or correlation) between successive observations of a single subject. Serial dependency seriously biases the traditional statistical tests (such as the *t* test or analysis of variance). Despite attempts to deal with serial dependency, the various proposed solutions are controversial and not yet widely accepted.

Single-subject designs have both advantages and disadvantages. Advantages include the ability to obtain functional relationships that apply to a single subject, the avoidance of artifacts that may emerge in group studies because of averaging data across subjects with differing behaviors, the potential identification of new important variables while attempting to stabilize baselines, and the ability to conduct experiments with an extremely limited number of available subjects. Disadvantages include the length of time required to test each subject through all the conditions of the experiment (increasing the possibility of subject loss because of attrition or equipment failure), the inability to detect the effects of weak variables when behavior is not well controlled, the difficulty in assessing the effects of variables that cause irreversible changes, and possibly limited external validity.

KEY TERMS

baseline design
behavioral baseline
stability criterion
baseline phase
intervention phase
ABAB design
intrasubject replication

reversal strategy
intersubject replication
systematic replication
direct replication
multiple-baseline design
dynamic design
discrete trials design

CHAPTER 13

Describing Data

CHAPTER OUTLINE

Descriptive Statistics and Exploratory Data Analysis
Organizing Your Data
Organizing Your Data for Computer Entry
Entering Your Data
Grouped Versus Individual Data
Graphing Your Data
Elements of a Graph
Bar Graphs
Line Graphs
Scatter Plots
Pie Graphs
The Importance of Graphing Data
The Frequency Distribution
Displaying Distributions
Examining Your Distribution
Descriptive Statistics: Measures of Center and Spread
Measures of Center
Measures of Spread
Boxplots and the Five-Number Summary
Measures of Association, Regression, and Related Topics
The Pearson Product-Moment Correlation Coefficient
The Point-Biserial Correlation
The Spearman Rank-Order Correlation
The Phi Coefficient
Linear Regression and Prediction
The Coefficient of Determination
The Correlation Matrix
Multivariate Correlational Techniques
Summary
Key Terms

Chapters 1 through 12 explored how to design and conduct research. Once you have conducted your research, the next step is to organize, summarize, and describe your data. This chapter reviews strategies that you can use to effectively organize, summarize, and describe data. The sections on *descriptive statistics* are intended to provide a brief review of these statistics. Computation is addressed in this chapter only where necessary to explain a particular statistic. If you need more information on these statistics, see one of the many introductory statistics texts available (such as Gravetter & Wallnau, 2010, or Pagano, 2010).

DESCRIPTIVE STATISTICS AND EXPLORATORY DATA ANALYSIS

Descriptive statistics allow you to summarize the properties of an entire distribution of scores with just a few numbers. Although descriptive statistics are commonly used to address the specific questions that you had in mind when you designed your study, they also can be employed to help you discover important but perhaps hidden patterns in your data that may shed additional light on the problems you are interested in resolving. The search for such patterns in your data is termed **exploratory data analysis (EDA)**. Over the past 40 years or so, a whole new set of descriptive tools has been developed to aid you in this search, many of which are graphical in nature.

When research has been designed to answer a specific question or set of questions, there is a strong temptation to rush directly to the inferential statistical techniques that will assess the "statistical significance" of the findings and to request only those descriptive statistics related directly to the analysis, such as group means, standard deviations, and standard errors. Resist this temptation. As we explain in Chapter 14, many of the most commonly used inferential statistics make certain crucial assumptions about the populations from which the scores in your data set were drawn. If these assumptions are violated, the results of the statistical analysis may be misleading.

Some exploratory techniques help you spot serious defects in your data that may warrant taking corrective action before you proceed to the inferential analysis. Others help you determine which summary statistics would be appropriate for a given set of data. Still others may reveal unsuspected influences. In this chapter, we introduce you to a number of descriptive tools (both numerical and graphical) for describing data and revealing secrets that hide within.

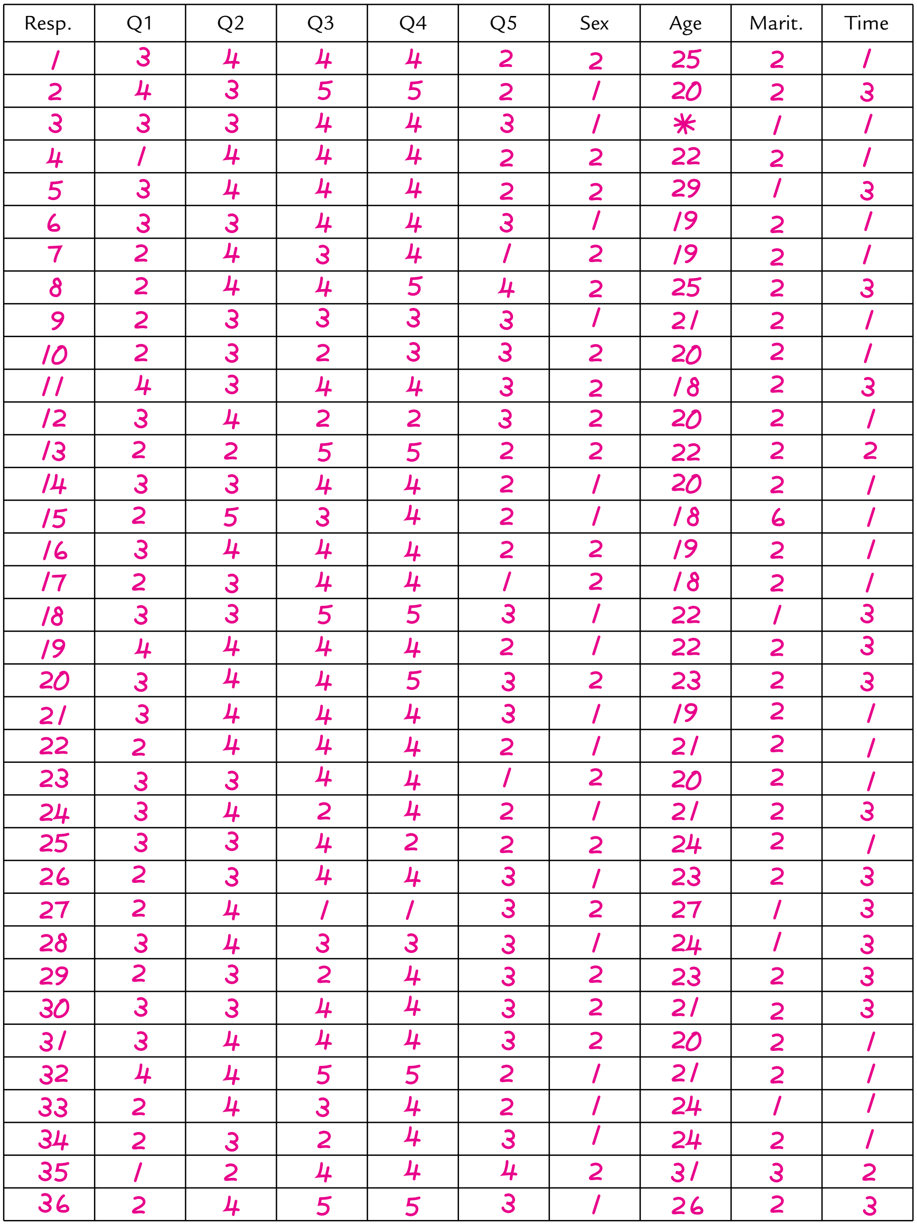
ORGANIZING YOUR DATA

Before you can interpret your data, you must first organize and summarize them. How you organize your data depends on your research design (whether you have conducted a survey, observational study, or experiment), how many variables were observed and recorded, and how observations were grouped or subdivided. A few representative examples follow.

For *survey data*, a data summary sheet like that shown in Figure 13-1 would be appropriate. The data are organized into a series of columns, one for numbering the respondents, one for each question asked, and one for each demographic item. To save space, the identifiers over the columns should be kept short. Here, we have simply labeled each question Q1, Q2, and so on. If space permits, you may invent more descriptive labels. Each row gives the data for one respondent. In the example, certain demographic variables have been dummy-coded. **Dummy codes** identify category values as numbers (e.g., sex of respondent: 1 = female, 2 = male). If the computer program that you are using accepts and can use category names, you need not dummy-code these variables. If it does not, you will have to use dummy coding.

Data sheets like the one shown in Figure 13-1 can go on for pages. A good strategy is to lay out your first data sheet and, before entering any data, copy it. In this way you avoid having to enter the column and row labels by hand on each new page. In addition, you should make out a single “key” or “code sheet” that describes the scale(s) used for the questions, gives a fuller description of the question or variable in each column, and indicates what each dummy code represents. Figure 13-2 shows the code sheet that accompanies the data sheet shown in Figure 13-1. The code sheet shows the Likert-scale codes used in conjunction with each attitude item, the five attitude statements, the three demographic items, and the dummy codes used to represent the category values of sex, marital status, and time of class attendance.

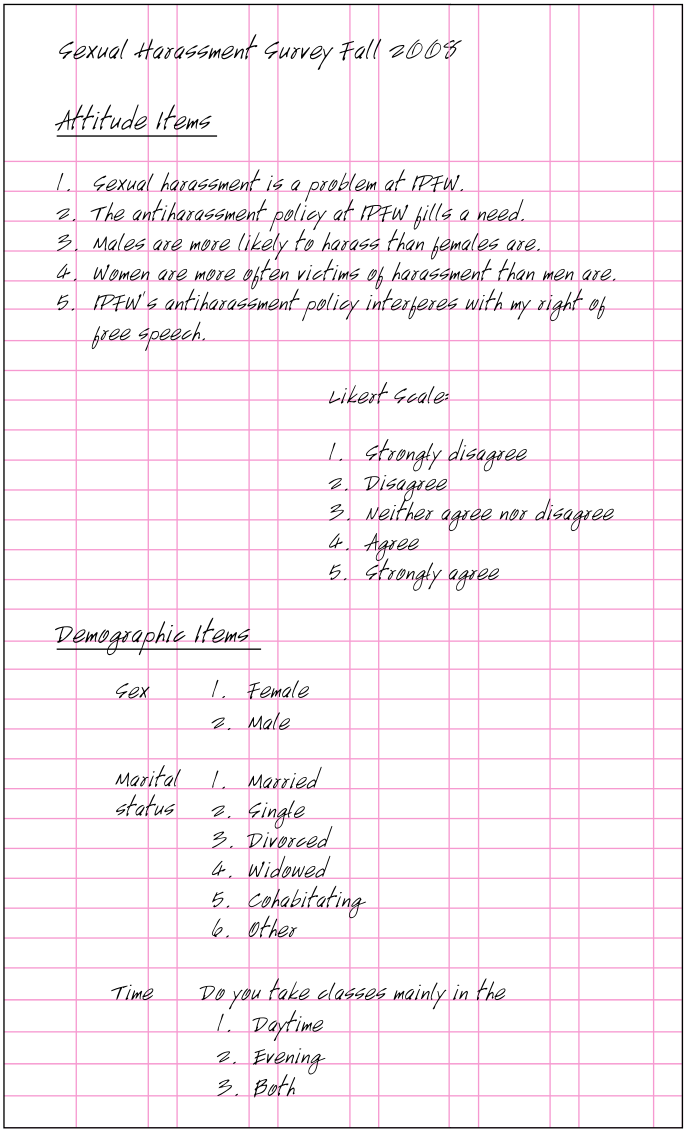
Experimental or quasi-experimental designs break down the dependent variable according to treatments or categories. You can organize data from these designs in two distinct ways. One way (called an *unstacked format*) is to create a separate column for the scores from each treatment. Figure 13-3 shows a simple summary sheet organized in this way for a 2 × 2 within-subjects factorial experiment. The subject numbers appear in the leftmost column. Because each subject was exposed to all the treatments, only one column of subject numbers was needed. Reserve space at the bottom of the data summary sheet for column summary statistics such as the mean and standard deviation. You can enter these after you have analyzed the data.



Resp.	Q1	Q2	Q3	Q4	Q5	Sex	Age	Marit.	Time
1	3	4	4	4	2	2	25	2	1
2	4	3	5	5	2	1	20	2	3
3	3	3	4	4	3	1	*	1	1
4	1	4	4	4	2	2	22	2	1
5	3	4	4	4	2	2	29	1	3
6	3	3	4	4	3	1	19	2	1
7	2	4	3	4	1	2	19	2	1
8	2	4	4	5	4	2	25	2	3
9	2	3	3	3	3	1	21	2	1
10	2	3	2	3	3	2	20	2	1
11	4	3	4	4	3	2	18	2	3
12	3	4	2	2	3	2	20	2	1
13	2	2	5	5	2	2	22	2	2
14	3	3	4	4	2	1	20	2	1
15	2	5	3	4	2	1	18	6	1
16	3	4	4	4	2	2	19	2	1
17	2	3	4	4	1	2	18	2	1
18	3	3	5	5	3	1	22	1	3
19	4	4	4	4	2	1	22	2	3
20	3	4	4	5	3	2	23	2	3
21	3	4	4	4	3	1	19	2	1
22	2	4	4	4	2	1	21	2	1
23	3	3	4	4	1	2	20	2	1
24	3	4	2	4	2	1	21	2	3
25	3	3	4	2	2	2	24	2	1
26	2	3	4	4	3	1	23	2	3
27	2	4	1	1	3	2	27	1	3
28	3	4	3	3	3	1	24	1	3
29	2	3	2	4	3	2	23	2	3
30	3	3	4	4	3	2	21	2	3
31	3	4	4	4	3	2	20	2	1
32	4	4	5	5	2	1	21	2	1
33	2	4	3	4	2	1	24	1	1
34	2	3	2	4	3	1	24	2	1
35	1	2	4	4	4	2	31	3	2
36	2	4	5	5	3	1	26	2	3

FIGURE 13-1 Example data summary sheet for survey data (* = missing data).

The second way to organize your data is to use a *stacked format*. In this format, you create one column for the participant IDs, a column for the treatment levels (dummy-coded), and a column for each dependent variable. Figure 13-4 redisplays a portion of the data of Figure 13-3 in this way. The stacked format works better than the unstacked format when your data include multiple independent



Sexual Harassment Survey Fall 2008

Attitude Items

1. Sexual harassment is a problem at IPFW.
2. The antiharassment policy at IPFW fills a need.
3. Males are more likely to harass than females are.
4. Women are more often victims of harassment than men are.
5. IPFW's antiharassment policy interferes with my right of free speech.

Likert Scale:

1. Strongly disagree
2. Disagree
3. Neither agree nor disagree
4. Agree
5. Strongly agree

Demographic Items

Sex
1. Female
2. Male

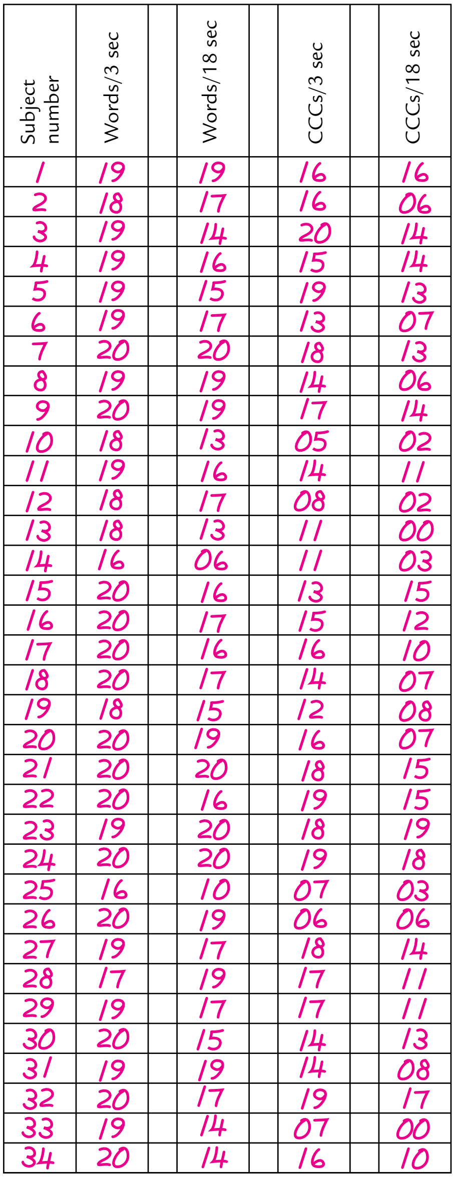
Marital status
1. Married
2. Single
3. Divorced
4. Widowed
5. Cohabitating
6. Other

Time
Do you take classes mainly in the
1. Daytime
2. Evening
3. Both

FIGURE 13-2 Code sheet for the data summary sheet of Figure 13-1.

or dependent variables. These are easily accommodated by including additional columns to indicate the treatment levels or observed values of the additional variables. Also, many computer statistical analysis packages expect the data to be entered in this format. A disadvantage of the stacked format is that, unlike the unstacked format, it does not provide a simple way to display treatment summary statistics.

More complex designs involving several independent and/or quasi-independent variables can be accommodated within either format. Figure 13-5 displays data in

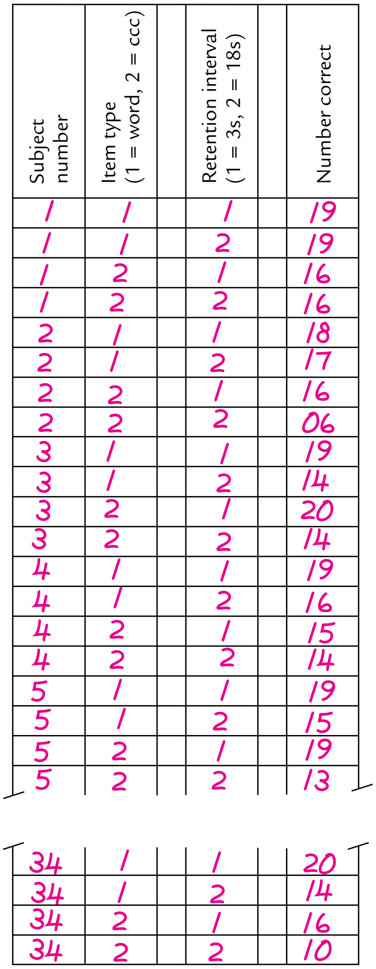


Subject number	Words/3 sec		Words/18 sec		CCCs/3 sec		CCCs/18 sec
1	19		19		16		16
2	18		17		16		06
3	19		14		20		14
4	19		16		15		14
5	19		15		19		13
6	19		17		13		07
7	20		20		18		13
8	19		19		14		06
9	20		19		17		14
10	18		13		05		02
11	19		16		14		11
12	18		17		08		02
13	18		13		11		00
14	16		06		11		03
15	20		16		13		15
16	20		17		15		12
17	20		16		16		10
18	20		17		14		07
19	18		15		12		08
20	20		19		16		07
21	20		20		18		15
22	20		16		19		15
23	19		20		18		19
24	20		20		19		18
25	16		10		07		03
26	20		19		06		06
27	19		17		18		14
28	17		19		17		11
29	19		17		17		11
30	20		15		14		13
31	19		19		14		08
32	20		17		19		17
33	19		14		07		00
34	20		14		16		10

FIGURE 13-3 Data summary sheet for a 2 × 2 within-subjects factorial experiment (unstacked format).

unstacked format from individual subjects for a 2 × 4 between-subjects design in a study by Bordens and Horowitz (1986) on the effects of joining multiple criminal offenses in a single trial (a procedure known as "joinder of offenses"). Each column provides the data from one treatment. In this two-factor between-subjects design, each treatment represented one level of "Charges judged" (the first independent variable) combined with one of the four levels of "Charges filed" (the second independent variable). The bottom two rows display summary measures (the mean and standard deviation) for each treatment.

FIGURE 13-4 Data summary sheet for the same data as shown in Figure 13-3, presented in stacked format.



Subject number	Item type (1 = word, 2 = ccc)		Retention interval (1 = 3s, 2 = 18s)		Number correct
1	1		1		19
1	1		2		19
1	2		1		16
1	2		2		16
2	1		1		18
2	1		2		17
2	2		1		16
2	2		2		06
3	1		1		19
3	1		2		14
3	2		1		20
3	2		2		14
4	1		1		19
4	1		2		16
4	2		1		15
4	2		2		14
5	1		1		19
5	1		2		15
5	2		1		19
5	2		2		13
34	1		1		20
34	1		2		14
34	2		1		16
34	2		2		10

A useful data summary sheet must be clearly labeled. Note that the columns of data in Figure 13-5 are clearly labeled with the levels of the independent variable in effect for each group. The top headings indicate the two levels of charges judged. The second level of headings indicates the level of charges filed as appropriate to each group.

The organization just described works well for a 2 × 4 factorial experiment and can be expanded to handle more levels of each factor or more factors. Other designs may require a different organization.

Organizing Your Data for Computer Entry

If you are going to submit your data to computer analysis, you should find out how the statistical analysis software that you intend to use expects the data to be organized. Many packages require the data to be entered in stacked format, some require the

	Charges judged							
	One charge				Two charges			
Charges filed	1	2	3	4	1	2	3	4
	3	5	6	2	6	3	4	5
	3	4	3	4	4	5	6	5
	3	4	4	5	4	5	4	5
	4	4	4	5	5	5	5	6
	4	5	5	5	4	4	5	5
	3	3	5	5	5	5	4	4
	2	3	4	5	5	4	4	5
	5	4	5	4	3	5	5	5
	6	4	3	6	3	5	5	6
	5	5	4	6	5	3	6	6
Mean	3.8	4.1	4.3	4.7	4.4	4.4	4.8	5.2
Standard deviation	1.23	0.74	0.95	1.16	0.97	0.84	0.79	0.63

FIGURE 13-5 Data summary sheet for a 2 × 4 between-subjects design.
SOURCE: Bordens and Horowitz, 1986.

unstacked format, and some will accept either. For example, a popular product called Minitab allows you to *enter* data in either format but may require one or the other, depending on the analysis requested. Certain commands allow you to change formats as required.

If you have not already done so on your original data sheets, you may have to code your variables before entering the data into the computer. Most software for data analysis looks for a numeric or alphabetic code to determine the levels or values of your independent and dependent variables. You must decide how to code these variables.

Coding independent variables involves assigning values to corresponding levels. For a quantitative independent variable (e.g., number of milligrams of a drug), simply record on your coding sheet the number of milligrams administered to subjects in each treatment group (e.g., 10, 20, or 30). For qualitative independent variables, you must assign an arbitrary number to each level. For example, if your independent variable were the loudness of a tone in an auditory discrimination experiment (low, moderate, and high), you might code the levels as 1 = low, 2 = moderate, 3 = high. As noted, this assignment of numbers to the levels of a qualitative independent variable is called *dummy coding*.

For quantitative data (e.g., if your participants rated the intensity of a sound on a scale ranging from 0 to 10), simply transfer each participant's score to your coding sheet. If, however, your dependent measure were qualitative (e.g., yes/no), you must dummy-code your dependent variable. For example, you could code all yes responses as 1 and all no responses as 2.

When coding your dependent variables, transfer the data (numeric or dummy-coded) to your coding sheet exactly as they are. Don't be concerned with creating new variables (e.g., by adding together existing ones) or with making special categories. Most statistical analysis software have commands that let you manipulate data in a variety of ways (adding numbers, doing data transformations such as a log transformation, etc.). So don't waste time creating new variables when preparing your data for input.

Entering Your Data

PC-based versions of statistical software have easy-to-use spreadsheet interfaces that allow you to enter data quickly and make corrections easily. If you don't like the data-entry provision in a particular program, the program may allow you to enter your data into a stand-alone spreadsheet program such as Microsoft's Excel and then read the spreadsheet file into the statistical program's data editor.

You can make data entry easier by organizing your data-coding sheet the way your data editor expects the data to be entered. For example, if your data will be entered one column at a time, organize the data into columns on the sheet. Then simply read down the columns while entering your data.

Errors climb when fatigue sets in, so if you are entering a large amount of data, take frequent breaks. Be sure to save any data that you have entered before you leave the keyboard.

Speaking of saving data, nothing can be more frustrating than spending half an hour entering data, only to have them obliterated by an unexpected power failure. You can minimize your losses on such occasions by frequently saving your data. (Most programs have a backup feature that you can configure to save the data automatically at periodic intervals.) Also, don't forget to save your data before you turn off the computer or exit from the data editor. Any data that you fail to save will be lost. At this time, you also should make a backup copy of your data on another disk or some other device such as a memory stick.

When you save your data, you create a *data file* from which the data will be read when the computer conducts your analysis. When you save the file, the computer will ask you to type in a file name under which the data will be saved. Try to think of a descriptive file name that will uniquely identify the data. When you have several data files, using descriptive file names makes it easier to find the correct file. It is also a good idea to add a date to your file name so that you will know which is the most recent version (e.g., MemoryApril10.dat).

After you have entered your data, check for errors. Because the computer cannot detect incorrectly entered data, it is up to you to catch any mistakes if you are to avoid invalid results. If you have had someone else enter the data for you, don't assume that the other person has already done the checking.

Software used to collect data via the Internet or on a computer may save the data in a format that you can directly read into a statistical analysis program. If this is the case, you will not have to go through the many steps of summarizing and transferring your data. Check the software's user manual or help files to see if this provision is available.

Grouped Versus Individual Data

After you have organized your data into a coherent format, you must decide on a basic descriptive strategy. In some cases, you may want to summarize your data by averaging scores for each group. In other cases, you may want to focus on the scores of individual subjects. Either strategy is valid, but each has its own advantages and disadvantages.

Grouped Data The major advantage of grouped data is convenience. When you calculate an average, you have one score that characterizes an entire distribution. You then can refer to the performance of subjects in a group by citing the average performance. If your data will be submitted to a statistical analysis based on treatment means, you will be treating your data in this way.

Although convenient, the grouped method does have two important limitations. First, the average score may not represent the performance of individual subjects in a group. An average score of 5 can result if all 10 subjects in a group scored 5 or if half scored 0 and half scored 10. In the former case, the average accurately reflects the individual performance of each subject. In the latter case, it does not. We examine this idea in more detail during the discussion of the mean.

The second limitation of using grouped data is that a curve resulting from plotting averaged data may not reflect the true nature of the psychological phenomenon being studied. In a learning experiment, for example, in which rats must meet a learning criterion (e.g., three consecutive error-free trials), a graph showing how the group average changes across trials might suggest that learning is a gradual process. Inspection of graphs of each individual subject's behavior might tell a different story. It might be that each rat evidences no learning for some variable number of trials, then suddenly masters the task, and after that never makes another error. Such a pattern of data would suggest that learning is an all-or-none proposition rather than the process of gradual improvement implied by the group average.

Individual Data Examining individual scores makes the most sense when you have repeated measures of the same behavior. Inspecting individual data also can be useful when the phenomenon under study is an either-or proposition (e.g., something was learned or not, or a stimulus was detected or not). In some cases, the individual data may reflect the effect of the independent variable more faithfully than data averaged over the group.

Using Grouped and Individual Data Together Researchers too often fall into the pattern of collecting data and then calculating an average without considering the individual scores constituting the average. A good strategy to adopt is to look at both the grouped and individual data. When you have repeated measures of the same behavior, examining individual data shows how each subject performed in your study. This may provide insights into the psychological process being studied that are not afforded by grouping data.

When you collect only a single score for each subject, you should still examine the distribution of individual scores. This usually entails plotting the individual scores on a graph and carefully inspecting the graph.

QUESTIONS TO PONDER

1. Why is it important to scrutinize your data using exploratory data analysis (EDA)?
2. How do you organize your data in preparation for data analysis?
3. What are the problems inherent in entering your data for computer data analysis?
4. What are the advantages and disadvantages of analyzing grouped and individual data?

GRAPHING YOUR DATA

Whether you have chosen a grouped or an individual strategy for dealing with your data, you will often find it beneficial to plot your data on a graph. Graphing helps you make sense of your data by representing them visually. The next sections describe the various types of graphs and indicate their uses. For details on drawing graphs, see Chapter 16.

Elements of a Graph

A basic graph represents your data in a two-dimensional space. The two dimensions (horizontal and vertical) are defined by two lines intersecting at right angles, called the *axes* of the graph. The horizontal axis is called the *abscissa* or *x-axis* of the graph, and the vertical axis is called the *ordinate* or *y-axis*. (The terms *x-axis* and *y-axis* are used in this discussion.)

When graphing data from an experiment, you normally represent levels of your independent variable along the *x*-axis and values of the dependent variable along the *y*-axis. A pair of values (one for the *x*-axis and one for the *y*-axis) defines a single *point* within the graph. You can present data within the two-dimensional space of a graph as a bar graph, line graph, scatter plot, or (abandoning the Cartesian *x*-axis, *y*-axis geometry) pie graph.

Bar Graphs

A **bar graph** presents your data as bars extending away from the axis representing your independent variable (usually the *x*-axis although this convention is not always followed). The length of each bar reflects the value of the dependent variable. Figure 13-6 shows group means from a one-factor, three-group experiment plotted as a bar graph. The three bars in Figure 13-6 represent the three levels of the independent variable for which data were collected. The length of each bar along the *y*-axis represents the mean score obtained on the dependent variable. Note that each bar straddles the *x*-axis value that it represents. The width of each bar has no meaning and is chosen to provide a pleasing appearance.

The bars usually represent estimates of population values based on sample data, such as the sample mean. In such cases the graph may also present an indication of

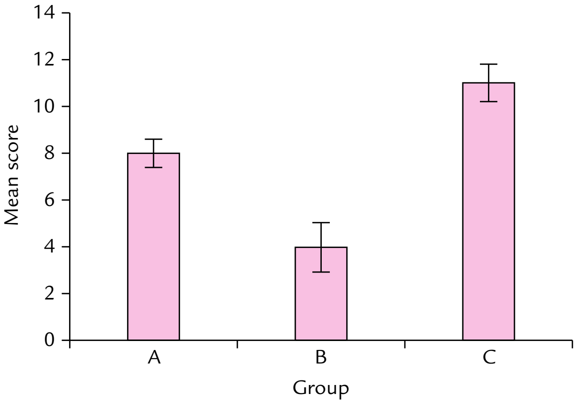


FIGURE 13-6 Bar graph from a hypothetical one-factor design, showing means and standard errors of the mean.

the precision of the estimate in the form of *error bars*, whiskers that extend from the tops of the main bars. The error bars show the variability of scores around the estimate. Figure 13-6 displays error bars depicting the standard error of the mean.

You also can use a bar graph to represent data from a multifactor design. Figure 13-7 shows a bar graph of the data from the two-factor joinder of offenses experiment (Bordens & Horowitz, 1986) described previously. Notice that the four levels of number of charges filed (one to four) are placed along the *x*-axis. The two levels of charges judged (the second independent variable) are represented within the graph itself. The gray bars represent the data from the one-charge judged group whereas the colored bars represent the data from the two-charges judged group.

A bar graph is the best method of graphing when your independent variable is categorical (such as the type of drug administered). In this case, the distance along the *x*-axis has no real meaning. A line graph (which visually emphasizes that distance) would be misleading. The bar graph makes the arbitrary ordering of categories apparent, whereas a line graph would inappropriately suggest the presence of trends in these data.

In addition to displaying such statistical values as treatment means, bar graphs may be used to display certain kinds of data distributions, discussed later in the chapter.

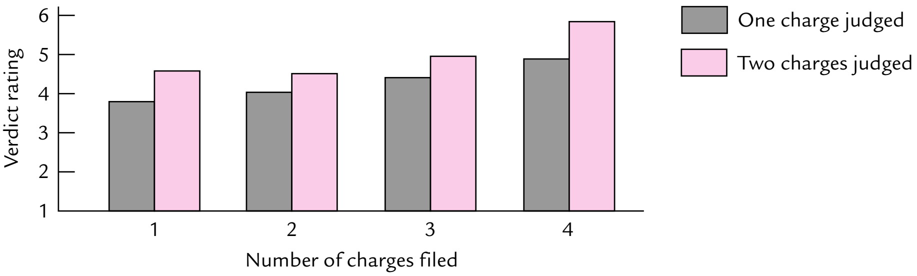


FIGURE 13-7 Bar graph of means from a two-factor design.

Line Graphs

A **line graph** represents data as a series of points connected by a line. It is most appropriate when your independent variable, represented on the *x*-axis, is continuous and *quantitative* (e.g., the number of seconds elapsing between learning and recall). This is in contrast to a bar graph, which is most appropriate when your independent variable is categorical or *qualitative* (e.g., categories representing grades on an exam). Line graphs are also appropriate when you want to illustrate functional relationships among variables. A *functional relationship* is one in which the value of the dependent variable varies as a function of the value of the independent variable. Usually, the depicted functional relationship is causal.

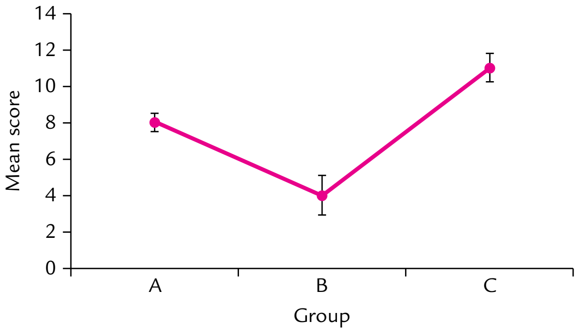
Figure 13-8 illustrates a line graph that depicts the group means from a single-factor experiment with a continuous independent variable. The error bars extending vertically in both directions depict the precision of the points as estimates of the population parameter, in this case represented by the standard error of the mean. These same data were shown in Figure 13-6 in the form of a bar graph. Notice the difference in how the two types of graphs visually represent the means.

A line graph also can be used to depict the means from multifactor experiments. Figure 13-9 shows such a line graph for the two-factor experiment on joinder of offenses (Bordens & Horowitz, 1986) described earlier. The levels of one factor are represented along the *x*-axis, just as in a single-factor experiment. The levels of the other factor are represented by using different symbols or line styles. All points collected under the same value of the second factor have the same symbol and are connected by the same line. Chapter 16 discusses how to draw line graphs.

Shapes of Line Graphs Relationships depicted on a line graph can take a variety of shapes. Figure 13-10 shows a graph on which the curve is *positively accelerated.* A positively accelerated curve is relatively flat at first and becomes progressively steeper as it moves along the *x*-axis. Positive acceleration can occur both in the upward and downward directions along the *y*-axis.

A curve also may be *negatively accelerated,* as shown in Figure 13-11. Here the curve is steep at first but becomes progressively flatter as it moves along the *x*-axis. Eventually, the curve "levels off" at some maximum or minimum value. The function is said to be *asymptotic* at this value. The *asymptote* of a curve is its theoretical limit,

FIGURE 13-8 Line graph showing means and standard errors from a one-factor design.



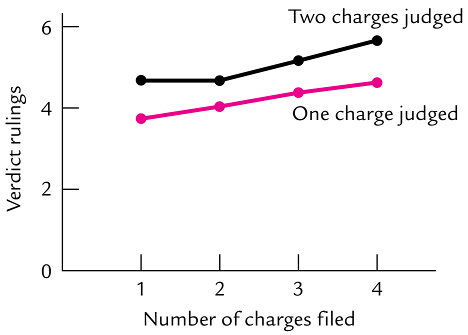


FIGURE 13-9 Line graph of means from a two-factor design.

or the point beyond which no further change in the value of the dependent variable is expected. In Figure 13-11 the relationship is asymptotic.

Whether positively or negatively accelerated, any curve also may be characterized as *increasing* or *decreasing*, which refers to whether the values along the *y*-axis increase or decrease, respectively, as the value along the *x*-axis increases. For example, a negatively accelerated, increasing function would approach a ceiling value at the asymptote whereas a negatively accelerated, decreasing function would approach a floor value.

A graph also may vary in complexity. The curves depicted in Figures 13-10 and 13-11 are both *monotonic*. That is, the curve represents a uniformly increasing or decreasing function. A *nonmonotonic* function contains reversals in direction, as illustrated in Figure 13-12. Notice how the curve changes direction twice by starting off low, rising, falling off, and then rising again.

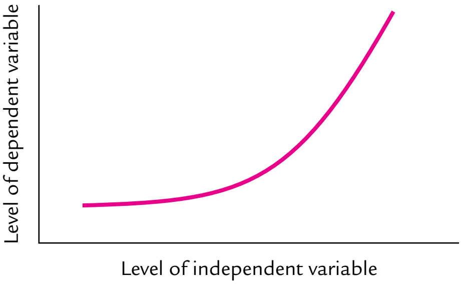


FIGURE 13-10 Line graph of positively accelerated functional relationship.

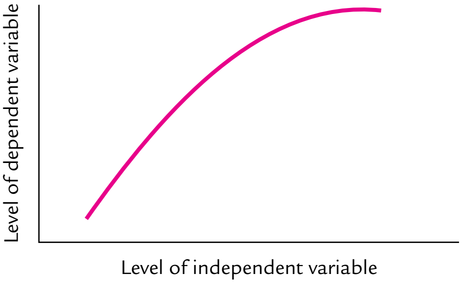
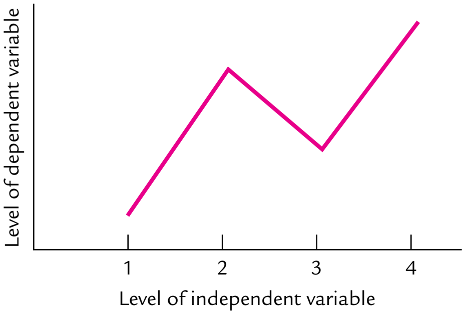


FIGURE 13-11 Line graph of negatively accelerated functional relationship.

FIGURE 13-12 Line graph of nonmonotonic functional relationship.



Scatter Plots

In research using a correlational strategy, the data from the two dependent measures are often plotted as a **scatter plot**. On a scatter plot, each pair of scores is represented as a point on the graph. For example, consider the data shown in Table 13-1. To make a scatter plot of these data, you plot the values of Variable A along the *x*-axis and the values of Variable B along the *y*-axis (or vice versa, it really does not matter). Then each pair of values is represented by a point within the graph. Figure 13-13 shows a scatter plot of the data in Table 13-1.

Scatter plots often include a "best-fitting" straight line (not shown in the figure) to indicate the general trend of the data points shown in the plot. In those cases the graph may also include the equation for this line and the coefficient of correlation. (We discuss these more fully in the section on correlation and regression below.)

Pie Graphs

If your data are in the form of proportions or percentages, then you might find a **pie graph** is a good way to represent the value of each category in the analysis. A pie graph represents the data as slices of a circular pie. Figure 13-14 shows two representative pie graphs. The pie graph to the left indicates the proportion of various behaviors observed in rat subjects during a half-hour coding period. The pie graph to the

TABLE 13-1 Bivariate Data for a Scatterplot

SUBJECT NUMBER	VARIABLE A	VARIABLE B
1	5	7
2	4	2
3	9	8
4	2	7
5	6	8
6	3	9

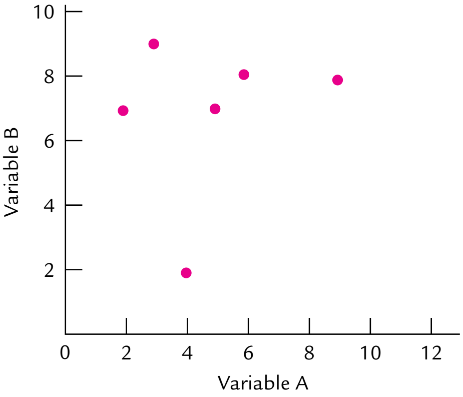


FIGURE 13-13 Scatter plot of the bivariate data presented in Table 13-1.

right, called an *exploded pie graph*, displays the same proportions while emphasizing the proportion of time devoted to grooming.

The Importance of Graphing Data

You can use either tables or graphs to summarize your data. If you organize data in tables, you present the numbers themselves (averages and/or raw score distributions). If you display the data in a graphical format, you lose some of this numerical precision. The value of a point usually can only be approximated by its position along the *y*-axis of the graph. However, graphing data is important for two major reasons, discussed in the next sections.

Showing Relationships Clearly The saying "One picture is worth a thousand words" applies to graphing data from your research. Although summarizing data in a table is fine, proper graphing adds a degree of clarity no table can provide. Consider the data presented in Table 13-2 and the same data graphically presented in Figure 13-9. Although both formats present the data accurately, the graph makes the relationships between the independent variables and dependent variable clearer. The graph brings out subtleties in the relationships that may not be apparent from inspecting a table.

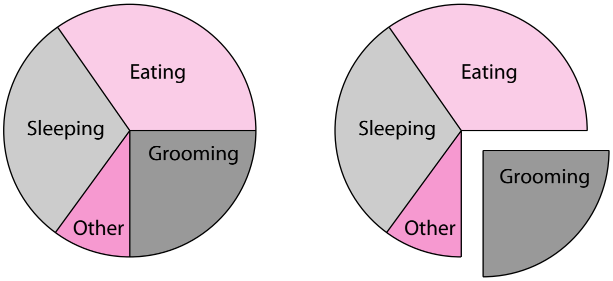


FIGURE 13-14 Pie graph and exploded pie graph.

TABLE 13-2 **Means From the 2 × 4 Joinder Experiment, in Tabular Format**

	NUMBER OF CHARGES FILED			
NUMBER OF CHARGES JUDGED	*One*	*Two*	*Three*	*Four*
One	3.8	4.1	4.3	4.7
Two	4.4	4.4	4.8	5.2

Choosing Appropriate Statistics In addition to making it easier to see relationships in your data, graphs allow you to evaluate your data for the application of an appropriate statistic. Before you apply any statistic to your data, graph your sample distributions and examine their shapes. Your choice of statistic will be affected by the manner in which scores are distributed, as described in the next section.

Graphing your data on a scatter plot is helpful when you intend to calculate a measure of correlation. Inspecting a scatter plot of your data can help you determine which measure of correlation is appropriate for your data. What you would look for and how your findings would affect your decision are taken up during the discussion of correlation measures later in the chapter.

THE FREQUENCY DISTRIBUTION

One of the first steps to perform when analyzing your data is to create a frequency distribution for each dependent variable in an experiment or for each variable in a correlational study. A **frequency distribution** consists of a set of mutually exclusive categories (*classes*) into which you sort the actual values observed in your data, together with a count of the number of data values falling into each category (*frequencies*). The classes may consist of response categories (e.g., for political party affiliation, they might consist of Democrat, Republican, Independent, and Other) or ranges of score values along a quantitative scale (e.g., for IQ, they might consist of 65–74, 75–84, 85–94, 95–104, 105–114, 115–124, and 125–134).

Displaying Distributions

Frequency distributions take the form of tables or graphs. Table 13-3 presents a hypothetical frequency distribution of IQ scores using the classes just given. Because IQ scores are quantitative data, the classes are presented in order of value from highest to lowest. To the right of each class is its frequency (*f*), the number of data values falling into that class. Because there were no IQ scores below 65 or above 134, classes beyond these limits are not tabled.

Although a table provides a compact summary of the distribution, it is not particularly easy to extract useful information from it about center, spread, and shape.

TABLE 13-3 Frequency Distribution Table of Hypothetical IQ Data

CLASS	*f*
125–134	5
115–124	12
105–114	22
95–104	25
85–94	26
75–84	7
65–74	3
Σ*f*	100

Graphical or semi-graphical displays are much better for this purpose. Here we describe two: the histogram and the stemplot.

The Histogram Figure 13-15 displays our IQ frequency distribution as a histogram. **Histograms** resemble bar graphs, with each bar representing a class. Unlike the bars in a bar graph, those in a histogram are drawn touching to indicate that there are no gaps between adjacent classes. Also, on a histogram, the *y*-axis represents a frequency: a count of the number of observations falling into a given category (e.g., the number of exam scores falling into the categories of A, B, C, D, or F). On a bar graph, the *y*-axis typically represents a mean score (e.g., the mean verdict rating shown in Figure 13-7).

The scale on which the variable was measured appears along the *x*-axis with the bars positioned appropriately to cover their respective ranges along the scale. The *y*-axis denotes the frequency; thus, a given bar's length indicates the frequency of scores falling within its range.

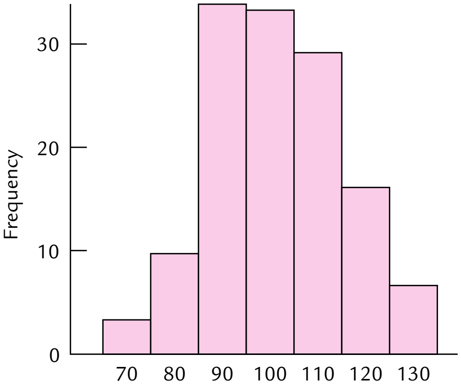


FIGURE 13-15 Hypothetical IQ data displayed as a histogram.

A histogram's appearance changes depending on how wide you make the classes. Make the classes too narrow, and you produce a flat-looking histogram with many empty or nearly empty classes. Make the classes too wide, and you produce a tall histogram lacking in detail. The goal is to create a histogram that shows reasonable detail without becoming flat and shapeless.

The Stemplot As a quick alternative to the histogram, you might consider using a **stemplot** (also known as a stem-and-leaf plot), which was invented by statistician John Tukey (1977), to simplify the job of displaying distributions. To create a stemplot of your data, you simply break each number into two parts: stem and leaf. The stem part might consist, for example, of the leftmost column or columns and the leaf part, the rightmost column. Thus, an IQ score of 67 would be broken into its leftmost number, or stem (6), and rightmost number, or leaf (7). After finding the lowest and highest stems, make a column that includes all the numbers in ascending order from lowest to highest stem. Then draw a vertical line immediately to the right of the stem column. Finally, for each score in your data, find its stem number and then write its leaf number on the same row immediately to the right of the stem. So, the IQ score of 67 would look like the first entry at the top of Figure 13-16. You do this for each number in your distribution. The final result would look something like Figure 13-16, which plots some hypothetical IQ data as a stemplot.

Stemplots are easy to construct and display and have the advantage over histograms and tables of preserving all the actual values present in the data. However, you do not have much freedom to choose the class widths because stemplots inherently create class widths of 10 (the span of a stem). Stemplots are not especially useful for larger data sets because the number of leaves becomes too large.

Examining Your Distribution

When examining a histogram or stemplot of your data, look for the following important features. First, locate the center of the distribution along the scale of measurement. In the IQ distribution plotted earlier, were the scores centered around 100 IQ points (an average value for the population as a whole) or somewhere else? The location of the center of a distribution tells you where the scores tended to cluster along the scale of measurement.

Second, note the spread of the scores. Do they tend to bunch up around the center or spread far from it? The spread of the scores indicates how variable they are.

Third, note the overall shape of the distribution. Is it hill shaped, with a single peak at the center, or does it have more than one peak? If hill shaped, is it more or

FIGURE 13-16 Hypothetical IQ data displayed as a stemplot.

Stem	Leaf
6	78
7	36
8	0222445555677789999
9	00000112344445566777889
10	011112222233445556666777889999
11	0233334455568889
12	0012557
13	02

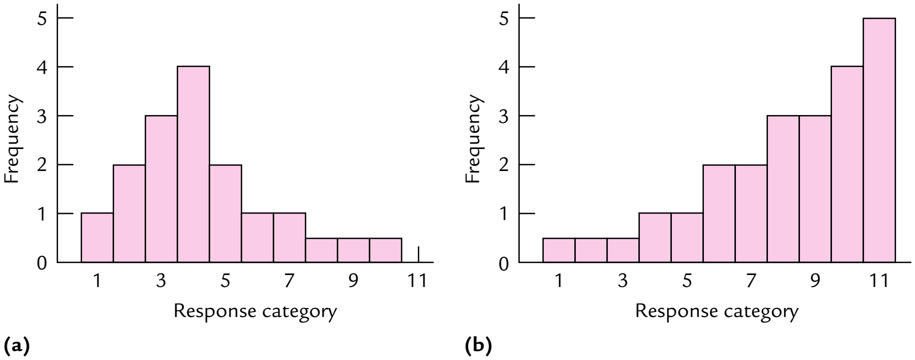


FIGURE 13-17 Two types of frequency distribution: (a) positively skewed and (b) negatively skewed.

less *symmetrical*, or is it skewed? A **skewed distribution** has a long "tail" trailing off in one direction and a short tail extending in the other. A distribution is *positively skewed* if the long tail goes off to the right, upscale (see Figure 13-17(a)) or *negatively skewed* if the long tail goes off to the left, downscale (see Figure 13-17(b)). Many variables encountered in psychology tend to produce a distribution that follows more or less a mathematical form known as the **normal distribution**, which is symmetric and hill shaped—the well-known bell curve. Because many common inferential statistics assume that the data follow a normal distribution, check the distribution of your data to see whether this assumption seems reasonable.

Finally, look for *gaps*, or **outliers**. Outliers are extreme scores that lie far from the others, well outside the overall pattern of the data (Moore & McCabe, 2006). Outliers may be perfectly valid (although unusual) scores, but sometimes they represent mistakes made in data collection or transcription. These bogus values can destroy the validity of your analysis. When you find an outlier, examine it carefully to determine whether it represents an error. Correct erroneous values or, if this is not possible, delete them from your analysis.

If you can find no valid reason for removing an outlier, you will have to live with it. However, you can minimize its effects on your analysis by using **resistant measures**, so called because they tend to resist distortion by outliers. We describe some of these measures next in our discussions of measures of center and spread.

QUESTIONS TO PONDER

1. How do various types of graphs differ, and when should each be used?
2. How do negatively accelerated, positively accelerated, and asymptotic functional relationships differ?
3. Why is it important to graph your data and inspect the graphs carefully?
4. How do you graph a frequency distribution as a histogram and as a stemplot?
5. What should you look for when examining the graph of a frequency distribution?

DESCRIPTIVE STATISTICS: MEASURES OF CENTER AND SPREAD

In many research situations, it is convenient to summarize your data by applying descriptive statistics. This section reviews two categories of descriptive statistics: measures of center and measures of spread. The next section describes another category of descriptive statistics, measures of association.

Measures of Center

A **measure of center** (also known as a *measure of central tendency*) gives you a single score that represents the general magnitude of scores in a distribution. This score characterizes your distribution by providing information about the score at or near the middle of the distribution. The most common measures of center are the mode, the median, and the mean (also called the *arithmetic average*). Each measure of center has strengths and weaknesses. Also, situations exist in which a given measure of center cannot be used.

The Mode The **mode** is simply the most frequent score in a distribution. To obtain the mode, count the number of scores falling into each response category. The response category with the highest frequency is the mode. The mode of the distribution 1, 2, 4, 6, 4, 3, 4 is 4.

No mode exists for a distribution in which all the scores are different. Some distributions, called *bimodal distributions*, have two modes. Figure 13-18 shows a bimodal distribution.

Although the mode is simple to calculate, it is limited because the values of scores outside of the most frequent score are not represented. The only information yielded by the mode is the most frequent score. The values of other data in the distribution are not taken into account. Under most conditions, take into account the other scores to get an accurate characterization of your data. To illustrate this point, consider the following two distributions of scores: 2, 2, 6, 3, 7, 2, 2, 5, 3, 1 and 2, 2, 21, 43, 78, 22, 33, 72, 12, 8.

In both these distributions, the mode is 2. Looking only at the mode, you might conclude that the two distributions are similar. Obviously, this conclusion is incorrect.

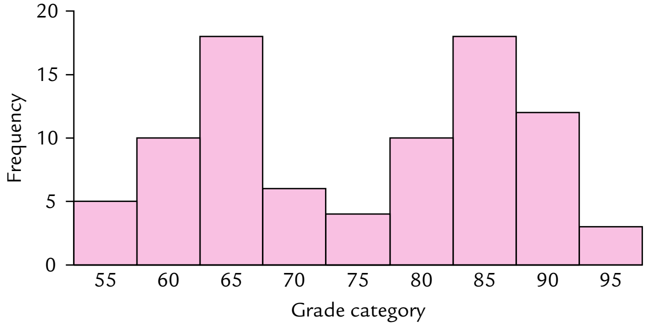


FIGURE 13-18 A bimodal frequency distribution.

It is clear that the second distribution is very different from the first. The mode may not represent a distribution very well and would not be the best measure to use when comparing distributions.

The Median A second measure of center is the median. The **median** is the middle score in an ordered distribution. To calculate the median, follow these steps:

1. Order the scores in your distribution from lowest to highest (or highest to lowest, it does not matter).
2. Count down through the distribution and find the score in the middle of the distribution. This score is the median of the distribution.

What is the median of the following distribution: 7, 5, 2, 9, 4, 8, 1? The correct answer is 5. The ordered distribution is 1, 2, 4, 5, 7, 8, 9, and 5 is the middle score.

You may be wondering what to do if you have an even number of scores in your distribution. In this case, there is no middle score. To calculate a median with an even number of scores, you order the distribution as before and then identify the *two* middle scores. The median is the average of these two scores. For example, with the ordered distribution of 1, 3, 6, 7, 8, 9, the median is 6.5 ($6 + 7 = 13$; $13/2 = 6.5$).

The median takes more information into account than the mode. However, it is still a rather insensitive measure of center because it does not take into account the magnitudes of the scores above and below the median. As with the mode, two distributions can have the same median and yet be very different in character. For this reason, the median is used primarily when the mean is not a good choice.

The Mean The **mean** (denoted as M) is the most sensitive measure of center because it takes into account all scores in a distribution when it is calculated. It is also the most widely used measure of center. The computational formula for the mean is

$$\mathrm{M} = \frac{\sum X}{n}$$

where $\sum X$ is the sum of the scores and n is the number of scores in the distribution. To obtain the mean, simply add together all the scores in the distribution and then divide by the total number of scores (n).

The major advantage of the mean is that, unlike the mode and the median, its value is directly affected by the magnitude of each score in the distribution. However, this sensitivity to individual score values also makes the mean susceptible to the influence of outliers. One or two such outliers may cause the mean to be artificially high or low. The following two distributions illustrate this point. Assume that Distribution A contains the scores 4, 6, 3, 8, 9, 2, 3, and Distribution B contains the scores 4, 6, 3, 8, 9, 2, 43. Although the two distributions differ by only a single score (3 versus 43), they differ greatly in their means (5 versus 10.7, respectively).

The mean of 5 appears to be more representative of the first distribution than the mean of 10.7 is of the second. The median is a better measure of center for the second distribution. The medians of the two distributions are 4 and 6, respectively—not nearly as different from one another as the means. Before you choose a measure of center, carefully evaluate your data for skewness and the presence of deviant, outlying scores. Do not blindly apply the mean just because it is the most sensitive measure of center.

Choosing a Measure of Center Which of the three measures of center you choose depends on two factors: the scale of measurement and the shape of the distribution of the scores. Before you use any measure of center, evaluate these two factors.

Chapter 5 described four measurement scales: nominal (qualitative categories), ordinal (rank orderings), interval (quantities measured from an arbitrary zero point), and ratio (quantities measured from a true zero point). The measurement scale that you chose when you designed your experiment will now influence your decision about which measure of center to use.

If your data were measured on a nominal scale, you are limited to using the mode. It makes no sense to calculate a median or mean sex, even if the sex of subjects has been coded as 0s (males) and 1s (females).

If your data were measured on an ordinal scale, you could properly use either the mode or the median, but it would be misleading to use the mean as your measure of center. This is because the mean is sensitive to the distance between scores. With an ordinal scale, the actual distance between points is unknown. You cannot assume that scores equally distant in terms of rank order are equally far apart, but you do assume this (in effect) if you use the mean.

The mean can be used if your data are scaled on an interval or ratio scale. On these two scales, the numerical distances between values are meaningful quantities.

Even if your dependent measure were scaled on an interval or ratio scale, the mean may be inappropriate. One of the first things you should do when summarizing your data is to generate a frequency distribution of the scores. Next, plot the frequency distribution as a histogram or stemplot and examine its shape. If your scores are normally distributed (or at least nearly normally distributed), then the mean, median, and mode will fall at the same point in the middle of the distribution, as shown in Figure 13-19. When your scores are normally distributed, use the mean as your measure of center because it is based on the most information.

As your distribution deviates from normality, the mean becomes a less representative measure of center. The two graphs in Figure 13-20 show the relationship between the three measures of center with a positively skewed distribution and a negatively

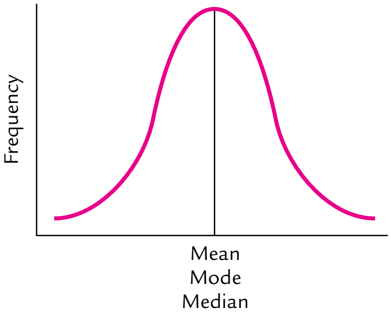


FIGURE 13-19 Line graph of normal distribution, showing location of mean, mode, and median.

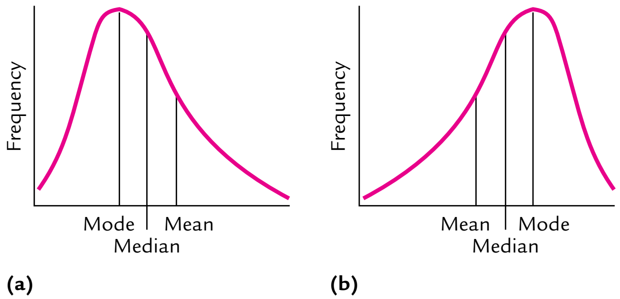


FIGURE 13-20 Line graph of (a) positively and (b) negatively skewed distributions, showing relationship between mean, mode, and median.

skewed distribution. Notice the relationship between the mean and median for these skewed distributions. In a negatively skewed distribution, the mean underestimates the center. Conversely, in a positively skewed distribution, the mean overestimates the center. Because the median is much less affected by skew, it provides a more representative picture of the distribution's center than does the mean and should be preferred whenever your distribution is strongly skewed.

Deviations from normality also create problems when deciding on an inferential statistic. Chapter 14 discusses inferential statistics and ways to deal with data that are not normally distributed. Neither the mean nor the median will accurately represent the center if your distribution is bimodal. With a bimodal distribution, both measures of center underrepresent one large cluster of scores and overrepresent the other.

Table 13-4 presents hypothetical scores from an introductory psychology exam that generated a bimodal distribution. These scores are shown graphically in Figure 13-18. The mean for these scores is 75.4, the median is 77, and both scores are in the grade C category. However, few students actually received a score in this range. The mean and median underrepresent the large cluster of scores in the grade B category and overestimate the large cluster of scores in the grade D category. Thus, neither the mean nor the median would be an appropriate measure of center for the scores in Table 13-4.

To summarize the discussion to this point, the three measures of center are the mean, the median, and the mode. The mean is the most sensitive measure of center because it takes into account the magnitude of each score in the distribution. The mean is also the preferred measure of center. The median is less sensitive to the distribution of scores than the mean but is preferred when your distribution is skewed or the distribution contains serious outliers. Which measure of center you can legitimately use depends on the scale on which the dependent variable was measured and on the manner in which the scores are distributed.

Measures of Spread

Another important descriptive statistic you should apply to your data is a **measure of spread** (also known as a *measure of variability*). If you look again at some of the sample distributions described thus far (or at the data presented in Table 13-1),

TABLE 13-4 **Hypothetical Scores on an Exam in an Introductory Psychology Class**

54	63	69	82	87
56	64	69	82	87
56	64	69	83	87
56	64	69	83	88
57	65	72	84	88
58	65	75	84	88
59	65	75	84	89
61	65	75	85	89
61	65	76	85	89
62	66	78	86	89
62	66	78	86	90
62	66	79	87	90
62	66	80	87	91
62	66	81	87	92
62	67	81	87	92
63	67	81	87	93
63	68	82	87	94

you will notice that the scores in the distributions differ from each other. When you conduct an experiment, it is extremely unlikely that your subjects will all produce the same score on your dependent measure. A measure of spread provides information that helps you to interpret your data. Two sets of scores may have highly similar means yet very different distributions, as the following example illustrates.

Imagine that you are a scout for a professional baseball team and are considering one of two players for your team. Each player has a .263 batting average over 4 years of college. The distributions of the two players' averages are as follows:

Player 1: .260, .397, .200, .195

Player 2: .263, .267, .259, .263

Which of these two players would you prefer to have on your team? Most likely, you would pick Player 2 because he is more "consistent" than Player 1. This simple example illustrates an important point about descriptive statistics. When you are evaluating your data, you should take into account both the center *and* the spread of the scores. This section reviews four measures of spread: the range, the interquartile range, the variance, and the standard deviation.

The Range The **range** is the simplest and least informative measure of spread. To calculate the range, you simply subtract the lowest score from the highest score. In the baseball example, the range for Player 1 is .202, and the range for Player 2 is .008.

Two problems with the range are that it does not take into account the magnitude of the scores between the extremes and that it is very sensitive to outliers in the distribution. Compare the following two distributions of scores: 1, 2, 3, 4, 5, 6 and 1, 2, 3, 4, 5, 31. The range for the first distribution is 5, and the range for the second is 30. The two ranges are highly discrepant despite the fact that the two distributions are nearly identical. For these reasons, the range is rarely used as a measure of spread.

The Interquartile Range The **interquartile range** is another measure of spread that is easy to calculate. To obtain the interquartile range, follow these steps:

1. Order the scores in your distribution.
2. Divide the distribution into four equal parts (quarters).
3. Find the score separating the lower 25% of the distribution (Quartile 1, or Q_1) and the score separating the top 25% from the rest of the distribution (Q_3). The interquartile range is equal to Q_3 minus Q_1.

The interquartile range is less sensitive than the range to the effects of extreme scores. It also takes into account more information because more than just the highest and lowest scores are used for its calculation. The interquartile range may be preferred over the range in situations in which you want a relatively simple, rough measure of spread that is resistant to the effects of skew and outliers.

The Variance The **variance** (s^2) is the average squared deviation from the mean. The defining formula is

$$s^2 = \frac{\sum(X - M)^2}{n - 1}$$

where X is each individual score making up the distribution, M is the mean of the distribution, and n is the number of scores. Table 13-5 shows how to use this formula by means of an example worked out for one distribution of scores.

The Standard Deviation Although the variance is frequently used as a measure of spread in certain statistical calculations, it does have the disadvantage of being expressed in units different from those of the summarized data. However, the variance can be easily converted into a measure of spread expressed in the *same* unit of measurement as the original scores: the **standard deviation** (s). To convert from the variance to the standard deviation, simply take the square root of the variance. The standard deviation of the data in Table 13-5 is 2.61 $\left(\sqrt{6.8}\right)$. The standard deviation is the most popular measure of spread.

Choosing a Measure of Spread The choice of a measure of center is affected by the distribution of the scores, and the same is true for the choice of a measure of spread. Like the mean, the range and standard deviation are sensitive to outliers. In cases in which your distribution has one or more outliers, the interquartile range may provide a better measure of spread.

TABLE 13-5 Calculation of a Variance

	X	X^2	$(X - M)$	$(X - M)^2$
	3	9	−2	4
	5	25	0	0
	2	4	−3	9
	7	49	2	4
	9	81	4	16
	4	16	−1	1
Σ	30	184		34
$M = 30/6 = 5.0$				
$s^2 = 34/5 = 6.8$				

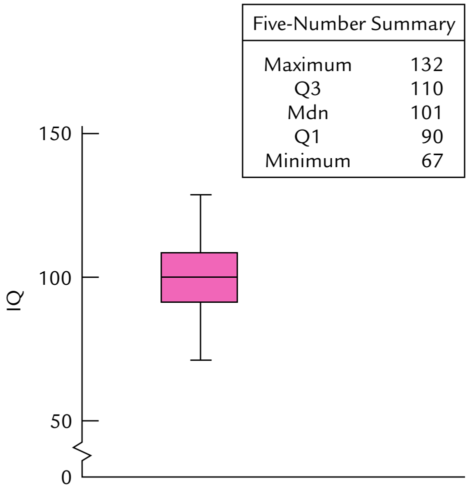
In addition to noting the presence of outliers, you should note the shape of the distribution (normal or skewed) when selecting a measure of spread. Remember that the mean is not a representative measure of center when your distribution of scores is skewed and that the mean is used to calculate the standard deviation. Consequently, with a skewed distribution, the standard deviation does not provide a representative measure of spread. If your distribution is seriously skewed, use the interquartile range instead.

Boxplots and the Five-Number Summary

The **five-number summary** provides a useful way to boil down a distribution into just a few easily grasped numbers, several of which are resistant to the effects of skew and outliers and all of which are based on the ranks of the scores. Included in the five-number summary are the following: the minimum, the first quartile, the median (second quartile), the third quartile, and the maximum. The minimum and maximum are simply the smallest and largest scores in the distribution; these are not resistant measures for the simple fact that the most extreme outliers will fall at the ends of the distribution and therefore are likely to *be* the maximum or minimum scores. The three center values (the first quartile, median, and third quartile) are resistant measures. From the five-number summary, you can easily calculate the range (maximum – minimum) and interquartile range ($Q_3 - Q_1$); the latter is of course a resistant measure of spread. By examining the five-number summary, you can quickly determine the center, spread, and range of the distribution in question.

An even better approach is to display the five-number summary as a **boxplot**. Figure 13-21 shows the five-number summary for our IQ distribution and displays these numbers as a boxplot. The first and third quartiles form the ends of the box, which encloses a line marking the median. The two "whiskers" reach out from the box to mark the minimum and maximum scores.

If you have data from several treatments or samples, you can easily compare the distributions from each using *side-by-side* boxplots, as shown in Figure 13-22. Each box should depict the distribution of the same variable.



Five-Number Summary	
Maximum	132
Q3	110
Mdn	101
Q1	90
Minimum	67

FIGURE 13-21 Five-number summary and boxplot of the IQ data.

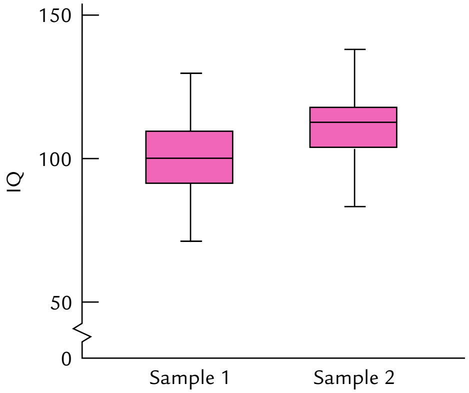


FIGURE 13-22 Side-by-side boxplots showing IQ data from two samples.

You often can discern the general shape of the distribution from the boxplot by noting the position of the median within the box and the relative lengths of the two whiskers. In a symmetric distribution, the median will fall close to the middle of the box, and the two whiskers will be similar in length. In a positively skewed distribution, the median will be pushed toward the left end or bottom of the box (nearer Q_1), and the right or top whisker will usually be longer than the left or bottom one. In a negatively skewed distribution, the reverse pattern will be found.

QUESTIONS TO PONDER

1. What is a measure of center?
2. How do the mode, median, and mean differ, and under what conditions would you use each?
3. What is a measure of spread?

4. What measures of spread are available, and when would you use each?
5. How are the variance and standard deviation related, and why is the standard deviation preferred?
6. What is the five-number summary, and how can you represent it graphically?

MEASURES OF ASSOCIATION, REGRESSION, AND RELATED TOPICS

In some cases, you may want to evaluate the direction and degree of relationship (correlation) between the scores in two distributions. For this purpose, you must use a *measure of association*. This section discusses several measures of association, along with the related topics of linear regression, the correlation matrix, and the coefficient of determination.

The Pearson Product-Moment Correlation Coefficient

The most widely used measure of association is the **Pearson product-moment correlation coefficient, or Pearson *r***. You would use it when you scale your dependent measures on an interval or a ratio scale. The Pearson correlation coefficient provides an index of the direction and magnitude of the relationship between two sets of scores.

The value of the Pearson *r* can range from +1 through 0 to −1. The sign of the coefficient tells you the direction of the relationship. A positive correlation indicates a *direct relationship* (as the values of the scores in one distribution increase, so do the values in the second). A negative correlation indicates an *inverse relationship* (as the value of one score increases, the value of the second decreases). Figure 13-23 illustrates scatter plots of data showing positive, negative, and no correlation.

The magnitude of the correlation coefficient tells you the degree of *linear relationship* (straight line) between your two variables. A correlation of 0 indicates that no relationship exists. As the strength of the relationship increases, the value of the correlation coefficient increases toward either +1 or −1. Both +1 and −1 indicate a perfect linear relationship. The sign is unrelated to the magnitude of the relationship and simply indicates the direction of the relationship. Figure 13-24 shows three correlations of differing strengths. Panel (a) shows a correlation of +1; panel (b), a correlation of about +.8; and panel (c), a correlation of 0.

Factors That Affect the Pearson r Before you use the Pearson *r*, examine your data much as you do when deciding on a measure of center. Several factors affect the magnitude and sign of the Pearson *r*.

The presence of outliers is one factor that affects the Pearson *r*. An outlier can drastically change your correlation coefficient and affect the magnitude of your correlation, its sign, or both. This is especially true if you use a small number of pairs of scores to compute the Pearson *r*.

Restricting the range over which the variables vary also can affect Pearson *r*. For example, if you were to examine the relationship between IQ and grade point average

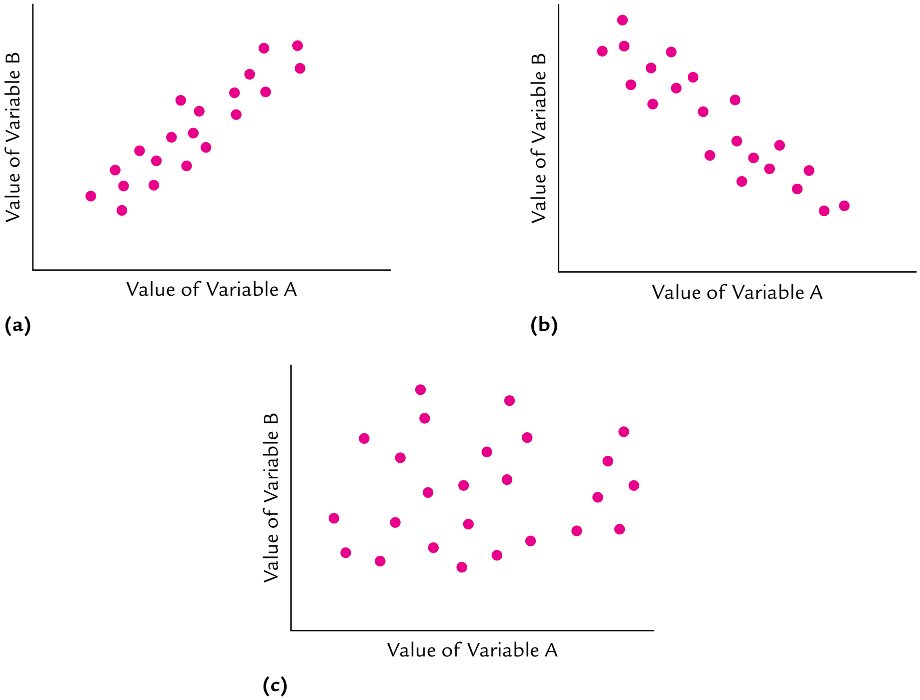


FIGURE 13-23 Scatter plots showing (a) positive, (b) negative, and (c) no correlation.

(GPA) in a group of college students, you would probably find a weaker correlation than if you examined the same two variables using high school students. Because IQ varies less among college students than among high school students, any variation in GPA that relates to IQ also will tend to vary less. As a result, the impact of extraneous variables such as motivation will be relatively larger, leading to a reduced correlation.

The Pearson *r* is sensitive to not only the range of the scores but also the shapes of the score distributions. The formula used to calculate the coefficient uses the standard deviation for each set of scores. Recall that you use the mean to calculate the standard deviation. If the scores are not normally distributed, the mean does not represent the distribution well. Consequently, the standard deviations will not accurately reflect the variability of the distributions, and the correlation coefficient will not provide an accurate index of the relationship between your two sets of scores. Hence, you should inspect the frequency distributions of each set of scores to ensure that they are normal (or nearly normal) before using the Pearson *r*.

Finally, the *Pearson r* reflects the degree to which the relationship between two variables is linear. Because of this assumption, take steps to determine whether the relationship appears to be linear. You can do this by constructing a scatter plot and then determining whether the points appear to scatter symmetrically around a straight line. Figure 13-25 shows a scatter plot in which the measures have a *curvilinear relationship* (rather than a linear relationship).

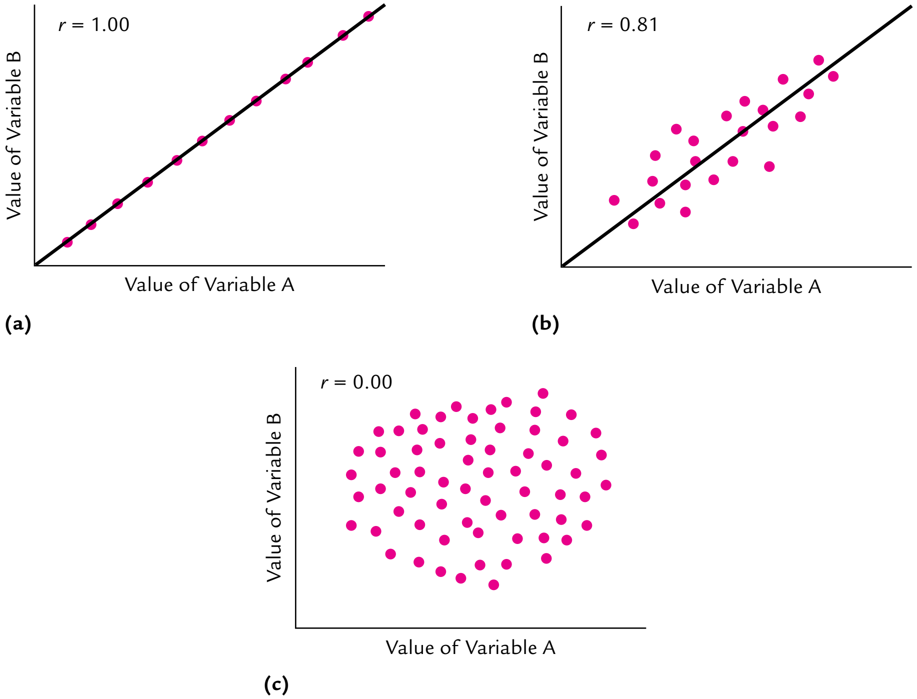


FIGURE 13-24 Scatter plots showing correlations of differing strengths: (a) perfect positive correlation, (b) strong positive correlation, and (c) zero correlation.

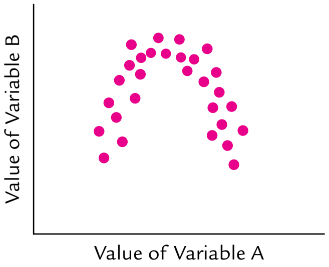


FIGURE 13-25 Scatter plot showing a curvilinear relationship.

When the relationship between variables is nonlinear, the Pearson *r* underestimates the degree of relationship between the variables. For example, the Pearson correlation between the variables illustrated in Figure 13-25 is zero. However, the two variables are obviously systematically related. There are special correlation techniques for nonlinear data, which are not discussed here.

The Pearson *r* is used when both of your variables are measured along a continuous scale. You may need to correlate variables when one (or both) of them is not measured along a continuous scale. Special correlation coefficients are designed for these purposes, three of which are discussed in the next sections.

The Point-Biserial Correlation

You may have one variable measured on an interval scale and the other measured on a nominal scale. For example, perhaps you want to investigate the relationship between self-rated political conservatism (measured on a 10-point scale) and whether or not a person voted for a particular referendum (yes or no). Because one variable is continuous and the other dichotomous (able to take on one of only two values), you would apply the **point-biserial correlation**.

Although there is a special formula for the point-biserial correlation, in practice you use the formula for the Pearson *r* to compute it. The dichotomous variable is dummy-coded as 0 for one response and 1 for the other. It is easier to use the Pearson formula, especially if you are using a computer program to evaluate your data (assuming the program cannot compute a point-biserial correlation).

Factors That Affect the Point-Biserial Correlation You should know a couple of things about the point-biserial correlation. First, its magnitude partly depends on the proportion of participants falling into each of the dichotomous categories. If the number of participants in each category is equal, then the maximum value the point-biserial can attain is ±1.0 (just as with the Pearson *r*). However, if the number of participants in each category is *not* equal, then the maximum attainable value for the point-biserial correlation is less than ±1.0. Consequently, the degree of relationship between the two variables may be underestimated. You should examine the proportion of participants using each category of the dichotomous variable and, if the proportions differ greatly, temper your conclusions accordingly.

The magnitude of the point-biserial correlation also is affected by the limited variation of the dichotomous variable (i.e., only two values possible). If the underlying variable is continuous but has been dichotomized for the analysis (e.g., anxiety level specified as either low or high), the point-biserial correlation will tend to underestimate the true strength of the relationship.

The Spearman Rank-Order Correlation

The **Spearman rank-order correlation**, or **rho** (ρ), is used either when your data are scaled on an ordinal scale (or greater) or when you want to determine whether the relationship between variables is monotonic (Gravetter & Wallnau, 2010). The rank-order correlation is relatively easy to calculate and can be interpreted in much the same way as a Pearson *r*.

The Phi Coefficient

The **phi coefficient (φ)** is used when *both* of the variables being correlated are measured on a dichotomous scale. You can calculate the phi coefficient with its own formula. However, like the point-biserial correlation, phi is usually calculated by dummy-coding the responses as 1s and 0s and then plugging the resulting scores into the formula for the Pearson *r*. The same arguments concerning restriction of range that apply to the point-biserial correlation also apply to phi—only doubly so.

QUESTIONS TO PONDER

1. What do measures of association tell you?
2. What are the measures of association available to you, and when would you use each?
3. What affects the magnitude and direction of a correlation coefficient?

Linear Regression and Prediction

A topic closely related to correlation is **linear regression**. With simple correlational techniques, you can establish the direction and degree of relationship between two variables. With linear regression, you can estimate values of a variable based on knowledge of the values of others. The following section introduces you to simple bivariate (two-variable) regression (also included are some calculations to help you understand regression). Chapter 15 extends bivariate regression to the case in which you want to consider multiple variables together in a single analysis.

Bivariate Regression The idea behind **bivariate linear regression** is to find the straight line that best fits the data plotted on a scatter plot. Consider an example using the data presented in Table 13-6, which shows the scores for each of 10 subjects on two measures (X and Y). Figure 13-26 shows a scatter plot of these data. You want to find the straight line that best describes the linear relationship between X and Y.

The best-fitting straight line is the one that minimizes the sum of the squared distances between each data point and the line, as measured along the *y*-axis (least-squares criterion). This line is called the **least-squares regression line**. At any given value for X found in the data, the position of the line indicates the value of Y predicted from the linear relationship between X and Y. You can then compare these predicted values with the values actually obtained. The best-fitting straight line minimizes the squared differences between the predicted and obtained values.

The following formula describes the regression line mathematically:

$$\hat{Y} = a + bX$$

where $\hat{Y}$ ("y-hat") is the predicted Y score, b is the slope of the regression line (also called the *regression weight*), X is the value of the X variable, and a is the *y*-intercept (Pagano, 2010). The constants a and b define a particular regression line. You can use the following formula to determine the value of b for a given set of data points (Gravetter & Wallnau, 2010):

$$b = \frac{SP}{SS_X}$$

where $SP = \Sigma\,(X - M_X)(Y - M_Y)$, $SS_X = \Sigma\,(X - M_X)^2$, and M_X and M_Y are the means for the X and Y scores, respectively.

TABLE 13-6 **Data for Linear Regression Example**

X	Y	$(X - M)$	$(Y - \bar{Y})$	$(X - M)(Y - \bar{Y})$	$(X - M)^2$
7	8	1.40	1.30	1.82	1.96
3	4	−2.60	−2.70	7.02	6.76
2	4	−3.60	−2.70	9.72	12.96
10	9	4.40	2.30	10.12	19.36
8	9	2.40	2.30	5.52	5.76
7	7	1.40	0.30	0.42	1.96
9	8	3.40	1.30	4.42	11.56
6	8	0.40	1.30	0.52	0.16
3	4	−2.60	−2.70	7.02	6.76
1	6	−4.60	−0.70	3.22	21.16
$M_X = 5.6$	$M_Y = 6.7$			$SP = 49.80$	$SS_X = 88.4$

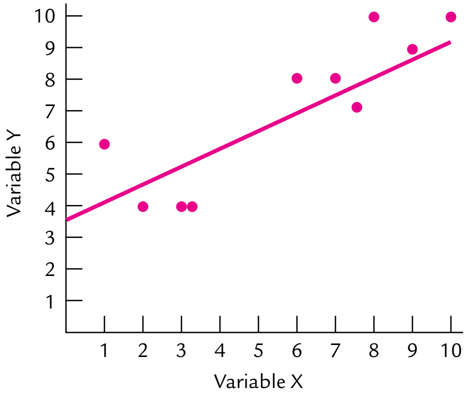


FIGURE 13-26 Scatter plot of data from Table 13-6.

Using the numbers from Table 13-6, we have

$$b = \frac{49.8}{88.4} = 0.56$$

The formula for the *y*-intercept (*a*) is

$$a = M_Y - bM_X$$

For this example,

$$a = 6.7 - 0.56(5.6) = 3.56$$

Substituting these values for *b* and *a* in the regression equation gives

$$\hat{Y} = 3.56 + 0.56X$$

This equation allows you to predict the value of Y for any given value of X. For example, if $X = 6$, then

$$\hat{Y} = 3.56 + 0.56(6) = 6.92$$

This regression equation was based on raw scores. Its **regression weight** (b) is known as a *raw score regression weight*. Raw score regression weights are difficult to interpret, so an alternative is normally used. If you plug standardized scores rather than raw scores into these equations, you will obtain a different regression equation with a different value for the weight and a zero value for the intercept. The regression weight that you obtain from this analysis is the *standardized regression weight*, or the *beta weight* (β). You use the standardized regression weights rather than the raw score regression weights when interpreting a regression equation. Chapter 15 discusses how to interpret standardized regression weights.

Residuals and Errors in Prediction After you have computed a regression analysis, you will have a score on one variable (Y) predicted from another variable (X). Because you have the actual values of a variable (Y), as well as the values predicted from the regression equation ($\hat{Y}$), you are in a position to see how accurately your regression equation predicts scores on Y. The difference between the values of Y and $\hat{Y}$ (i.e., $Y - \hat{Y}$) is a *residual*. Residuals will be low when the regression equation generates values of $\hat{Y}$ that are close to the actual values of Y.

Perfectly correlated variables result in no error in prediction (the predicted and actual values of Y will always agree). However, when your correlation is less than perfect, there will be error in predicting Y from X. You can estimate the amount of error in prediction by calculating the **standard error of estimate**, which is a measure of the distance between your data points and your computed regression line (Gravetter & Wallnau, 2010). The following formula is used to compute the standard error of estimate (Gravetter & Wallnau, 2010; *df* stands for "degrees of freedom"):

$$s_{est} = \sqrt{\frac{SS_{error}}{df}} = \sqrt{\frac{\Sigma(Y - \hat{Y})}{n - 2}}$$

In the current example, $s_{\text{est}} = 1.008$.

A close but inverse relationship exists between the magnitude of s_{est} and the magnitude of the correlation between X and Y. If X and Y are highly correlated, the data points will be clustered tightly around the regression line, and s_{est} will be small. As the strength of the relationship between X and Y decreases, s_{est} increases.

The Coefficient of Determination

The square of the correlation coefficient (whether Pearson r, point-biserial, Spearman rho, or phi) is called the *coefficient of determination*. The coefficient of determination provides a measure of the amount of variance that two variables being tested share. It indicates how much of the variability in one of the scores can be "explained" by the variability in the other score. For example, if variation in Score X actually *caused* variations to occur in Score *Y*, the coefficient of determination would indicate what proportion of the total variation in Score *Y* was caused by variation in Score *X*.

As an example, assume that you investigated the relationship between intelligence and school performance and found a correlation of .60. Then the coefficient of determination is .60 × .60, or .36. This means that 36% of the variation in school performance is accounted for by the variation in intelligence.

Of course, you usually don't know if the relationship is truly a causal one or in which direction the causal arrow points. Consequently, you should interpret this statistic with caution. Perhaps the most enlightening use of this statistic is to subtract it from 1.0. The resulting number, called the **coefficient of nondetermination**, gives the proportion of variance in one variable *not accounted for* by variance in the other variable. This is in effect *unexplained* variance caused by unmeasured factors. If the coefficient of nondetermination is large, then your measured variables are having little impact on each other relative to these unmeasured factors. If this happens, then perhaps you should try to identify these unmeasured variables and either hold them constant or measure them.

QUESTIONS TO PONDER

1. What is linear regression, and how is it used to analyze data?
2. How are regression weights and standard error used to interpret the results from a regression analysis?
3. What is the coefficient of determination and what does it tell you?
4. What it the coefficient of nondetermination, and what does it tell you?

The Correlation Matrix

If you have computed all the possible correlations among a number of variables, you can make the relationships among the variables easier to comprehend by displaying the correlation coefficients in a table called a **correlation matrix**. Table 13-7 shows a hypothetical correlation matrix for five variables (1 to 5). The variables being correlated in the matrix are shown in the headings along the top and left side of the matrix. Each number within the matrix is the correlation between the two variables whose row and column intersect at the position of the number. For example, the correlation between Variables 5 and 3 can be found by reading across the row labeled "Variable 5" to the column labeled "Variable 3." The correlation found at that intersection is .06.

Note that the numbers along the diagonal are omitted from the table. This is because the diagonal positions represent the correlations of each variable with itself, which are necessarily 1.0. You also omit the correlations above the diagonal because they simply duplicate the correlations already given below the diagonal. For example, the correlation of Variable 5 with Variable 3 (below the diagonal) is the same as the correlation of Variable 3 with Variable 5 (which would appear above the diagonal).

TABLE 13-7 A Correlation Matrix

	VARIABLES			
VARIABLES	*1*	*2*	*3*	*4*
2	.54			
3	.43	.87		
4	.52	.31	.88	
5	.77	.44	.06	.39

Multivariate Correlational Techniques

The measures of correlation and linear regression discussed in this chapter are all bivariate. Even if you calculate several bivariate correlations and arrange them in a matrix, your conclusions are limited to the relationship between pairs of variables. Bivariate correlation techniques are certainly useful and powerful tools. In many cases, however, you may want to look at three or more variables simultaneously. For example, you might want to know what the relationship between two variables is with the effect of a third held constant. Or you might want to know how a set of predictor variables relates to a criterion variable. In these cases and related others, the statistical technique of choice is *multivariate analysis*. Multivariate analysis is a family of statistical techniques that allow you to evaluate complex relationships among three or more variables. Multivariate analyses include multiple regression, discriminant analysis, part and partial correlation, and canonical correlation. Chapter 15 provides an overview of these and other multivariate techniques.

QUESTIONS TO PONDER

1. What is a correlation matrix, and why should you construct and inspect one?
2. How does a multivariate correlational statistic differ from a bivariate correlational statistic?

SUMMARY

When you have finished conducting your research, you begin the task of organizing, summarizing, and describing your data. The first step is to organize your data so that you can more easily conduct the relevant analyses. A good way to gain some understanding of your data is to graph the observed relationships. You can do this with a bar graph, line graph, scatter plot, or pie graph, whichever is most appropriate for your data.

A frequency distribution shows how the scores in your data vary along the scale of measurement. Although a frequency distribution can be presented in tabular format, you can grasp its essential features more easily by graphing it as a histogram or by creating a stemplot. When examining these, you should look for several important features: the center, around which the scores tend to vary; the spread, or degree to which the scores tend to vary from the center; the overall shape of the distribution (e.g., symmetric or skewed); and the presence of gaps or outliers—deviant points lying far from the rest. Examine any outliers carefully and correct or eliminate any that resulted from error. Descriptive statistics are methods for summarizing your data. Descriptive statistics include measures of central tendency, measures of variability, and measures of correlation.

The mode, median, and mean are the three measures of center. The mode is the most frequent score in your distribution. The median is the middle score in an ordered distribution. The mean is the arithmetic average of the scores, obtained by summing the scores and dividing the sum by the total number of scores.

Which of the three measures of center that you should use depends both on the scale that the data were measured on, and on the shape of the distribution of scores. The mean can be used only with data that are scaled on either a ratio or an interval scale and are normally distributed. In cases in which the data are skewed or bimodal, then the mean does not provide a representative measure of center, and the median or mode should be considered. Ordinally scaled data are best described with the median, and nominally scaled data are best described with the mode.

Measures of spread include the range, interquartile range, variance, and standard deviation. The range is simply the difference between the highest and lowest scores in your distribution. Although simple to calculate, the range is rarely used. Serious limitations of the range are that it is strongly affected by extreme scores and takes into account only the highest and lowest scores (thus ignoring the remaining scores in the distribution). The interquartile range takes into account more of the scores in the distribution and is less sensitive than the range to extreme scores. The variance uses all the scores in its calculation but has the disadvantage that its unit of measurement differs from that of the scores from which it derives. This problem can be overcome by taking the square root of the variance. The resulting statistic, the standard deviation, is the most commonly used measure of spread.

Your decision about which of the measures of spread to use is affected by the same two factors that affect your decision about central tendency (scale of measurement and distribution of scores). The standard deviation is a good measure of spread when your scores are normally distributed. As scores deviate from normality, the standard deviation becomes a less representative measure of spread. When your data are skewed, use the interquartile range.

The five-number summary provides a concise view of your distribution by providing the minimum, first quartile, median, third quartile, and maximum. Displaying these five numbers as a boxplot helps visualize the center, spread, and shape of the distribution. You can quickly compare distributions of the same variable from different treatments or samples by creating side-by-side boxplots.

Measures of correlation provide an index of the direction and degree of relationship between two variables. The most popular measure of correlation is the Pearson product-moment correlation coefficient (r). This coefficient can range from -1 through 0 to $+1$. A stronger relationship is indicated as the coefficient approaches ± 1. A negative correlation indicates that an increase in the value of one variable is associated with a decrease in the value of the second (inverse relationship). A positive correlation indicates that the two measures increase or decrease together (direct relationship).

The Pearson r is applied to data scaled on either an interval or a ratio scale. Other measures of correlation are available for data measured along other scales. The point-biserial correlation is used if one variable is measured on an interval or ratio scale and the other on a dichotomous nominal scale. Spearman's rho is used if both variables are measured on at least an ordinal scale. The phi coefficient is used if both variables are dichotomous.

Linear regression is a statistical procedure closely related to correlation. With linear regression, you can estimate the value of a criterion variable given the value of a predictor. In linear regression, you calculate a least-squares regression line, which is the straight line that best fits the data on a scatter plot. This line minimizes the sum of the squared distances between each data point and the line, as measured along the y-axis (least-squares criterion), and minimizes the difference between predicted and obtained values of y. The amount of discrepancy between the values of y predicted with the regression equation and the actual values is provided by the standard error of estimate. The magnitude of the standard error is related to the magnitude of the correlation between your variables. The higher the correlation, the lower the standard error.

By squaring the correlation coefficient, you obtain the coefficient of determination, an index of the amount of variation in one variable that can be accounted for by variation in the other. Subtracting the coefficient of determination from 1.0 gives you the coefficient of nondetermination, the proportion of variance *not* shared by the two variables. The larger this number, the larger the effect of unmeasured sources of variance relative to that of the measured variables.

Multivariate statistical techniques are used to evaluate more complex relationships than simple bivariate statistics. With multivariate statistics, you can analyze the degree of relationship between a set of predictor variables and a criterion variable or look at the correlation between two variables with the effect of a third variable held constant.

KEY TERMS

descriptive statistics
exploratory data analysis (EDA)
dummy code
bar graph
line graph
scatter plot
pie graph
frequency distribution
histogram
stemplot
skewed distribution
normal distribution
outlier
resistant measure
measure of center
mode

median

mean

measure of spread

range

interquartile range

variance

standard deviation

five-number summary

boxplot

Pearson product-moment correlation coefficient, or Pearson *r*

point-biserial correlation

Spearman rank-order correlation (rho)

phi coefficient (φ)

linear regression

bivariate linear regression

least-squares regression line

regression weight

standard error of estimate

coefficient of nondetermination

correlation matrix

14 CHAPTER

Using Inferential Statistics

CHAPTER OUTLINE

Inferential Statistics: Basic Concepts
- Sampling Distribution
- Sampling Error
- Degrees of Freedom
- Parametric versus Nonparametric Statistics

The Logic Behind Inferential Statistics
- Statistical Errors
- Statistical Significance
- One-Tailed Versus Two-Tailed Tests

Parametric Statistics
- Assumptions Underlying a Parametric Statistic
- Inferential Statistics with Two Samples
- The *t* Test
- An Example from the Literature: Contrasting Two Groups
- The *z* Test for the Difference Between Two Proportions
- Beyond Two Groups: Analysis of Variance (ANOVA)
- The One-Factor Between-Subjects ANOVA
- The One-Factor Within-Subjects ANOVA
- The Two-Factor Between-Subjects ANOVA
- The Two-Factor Within-Subjects ANOVA
- Mixed Designs
- Higher-Order and Special-Case ANOVAs
- ANOVA: Summing Up

Nonparametric Statistics
- Chi-Square
- The Mann–Whitney *U* Test
- The Wilcoxon Signed Ranks Test
- Parametric Versus Nonparametric Statistics

Special Topics in Inferential Statistics
- Power of a Statistical Test
- Statistical Versus Practical Significance
- The Meaning of the Level of Significance
- Data Transformations
- Alternatives to Inferential Statistics

Summary

Key Terms

Chapter 13 reviewed descriptive statistics that help you characterize and describe your data. However, they do not help you assess the reliability of your findings. A reliable finding is repeatable whereas an unreliable one may not be. Statistics that assess the reliability of your findings are called **inferential statistics** because they let you infer the characteristics of a population from the characteristics of the samples comprising your data.

This chapter reviews the most widely used inferential statistics. Rather than focusing on how to calculate these statistics, this discussion focuses on issues of application and interpretation. Consequently, computational formulas or worked examples are not presented.

INFERENTIAL STATISTICS: BASIC CONCEPTS

Before exploring some of the more popular inferential statistics, we present some of the basic concepts underlying these statistics. You should understand these concepts before tackling the discussion on inferential statistics that follows. If you need a more comprehensive refresher on these concepts, consult a good introductory statistics text.

Sampling Distribution

Chapter 13 introduced the notion of a distribution of scores. Such a distribution results from collecting data across a series of observations and then plotting the frequency of each score or range of scores. It is also possible to create a distribution by repeatedly taking samples of a given size (e.g., $n = 10$ scores) from the population. The means of these samples could be used to form a distribution of sample means. If you could take *every possible* sample of n scores from the population, you would have what is known as the *sampling distribution of the mean*. Statistical theory reveals that this distribution will tend to closely approximate the normal distribution, even when the population of scores from which the samples were drawn is far from normal in shape. Thus, you can use the normal distribution as a theoretical model that will allow you to make inferences about the likely value of the population mean, given the mean of a single sample from that population.

The sample mean is not the only statistic for which you can obtain a sampling distribution. In fact, each sample statistic has its own theoretical sampling distribution. For example, the tabled values for the *z* statistic, Student's *t*, the *F* ratio, and chi-square represent the sampling distributions of those statistics. Using these sampling distributions, you can determine the probability that a value of a statistic as large as or larger than the obtained value would have been obtained if only chance were at work. This probability is called the obtained *p*.

Sampling Error

When you draw a sample from a population of scores, the mean of the sample, M, will probably differ from the population mean, μ. An estimate of the amount of variability in the expected sample means across a series of such samples is provided by the **standard error of the mean** (or *standard error* for short). It may be calculated from the standard deviation of the sample as follows:

$$s_{\mathrm{M}} = \frac{s}{\sqrt{n}}$$

where *s* is the standard deviation of the sample and *n* is the number of scores in the sample. The standard error is used to estimate the standard deviation of the sampling distribution of the mean for the population from which the sample was drawn.

Degrees of Freedom

In any distribution of scores with a known mean, a limited number of data points yield independent information. For example, if you have a sample of 10 scores and a known mean (e.g., 6.5), only 9 scores are free to vary. That is, once you have selected 9 scores from the population, the value of the 10th must have a particular value that will yield the mean. Thus, the **degrees of freedom (*df*)** for a single sample are $n - 1$ (where *n* is the total number of scores in the sample).

Degrees of freedom come into play when you use any inferential statistic. You can extend this logic to the analysis of an experiment. If you have three groups in your experiment with means of 2, 5, and 10, the grand mean (the sum of all the scores divided by *n*) is then 5.7. If you know the grand mean and you know the means from two of your groups, the final mean is set. Hence, the degrees of freedom for a three-group experiment are $k - 1$ (where *k* is the number of levels of the independent variable). The degrees of freedom are then used to find the appropriate tabled value of a statistic against which the computed value is compared.

Parametric Versus Nonparametric Statistics

Inferential statistics can be classified as either *parametric* or *nonparametric*. A *parameter* in this context is a characteristic of a population, whereas a *statistic* is a characteristic of your sample (Gravetter & Wallnau, 2010). A *parametric statistic* estimates the value of a population parameter from the characteristics of a sample.

When you use a parametric statistic, you are making certain assumptions about the population from which your sample was drawn. A key assumption of a parametric test is that your sample was drawn from a normally distributed population.

In contrast to a parametric statistic, a *nonparametric statistic* makes no assumptions about the distribution of scores underlying your sample. Nonparametric statistics are used if your data do not meet the assumptions of a parametric test.

QUESTIONS TO PONDER

1. Why are sampling distributions important in inferential statistics?
2. What is sampling error, and why is it important to know about?
3. What are degrees of freedom, and how do they relate to inferential statistics?
4. How do parametric and nonparametric statistics differ?

THE LOGIC BEHIND INFERENTIAL STATISTICS

Whenever you conduct an experiment, you expose subjects to different levels of your independent variable. Although a given experiment may contain several groups, assume for the present discussion that the experiment in question includes only two. The data from each group can be viewed as a sample of the scores obtained if all subjects in the target population were tested under the conditions to which the group was exposed. For example, the treatment group mean represents a population of subjects exposed to your experimental treatment. Each treatment mean is assumed to represent the mean of the underlying population.

In all respects except for treatment, the treatment and control groups were exposed to equivalent conditions. Assume that the treatment had *no effect* on the scores. In that case, each group's scores could be viewed as an independent sample taken from the *same* population. Figure 14-1 illustrates this situation.

Each sample mean provides an independent estimate of the population mean. Each sample standard error provides an independent estimate of the standard deviation of sample means in the sampling distribution of means. Because the two means were drawn from the same population, you would expect them to differ only because of sampling error. You can assume that the distribution of these means is normal (central limit theorem), and you have two estimates of the standard deviation of this distribution (the standard errors). From this information, you can calculate the probability that the two sample means would differ as much as or more than they do simply because of chance factors. This probability is the obtained p.

Let's review these points. If the treatment had no effect on the scores, then you would expect the scores from the two groups to provide independent samples from the same population. From these samples, you can estimate the characteristics of that population; from this estimate, you can determine the probability that sampling error and sampling error alone would produce a difference at least as large as the observed difference between the two treatment means.

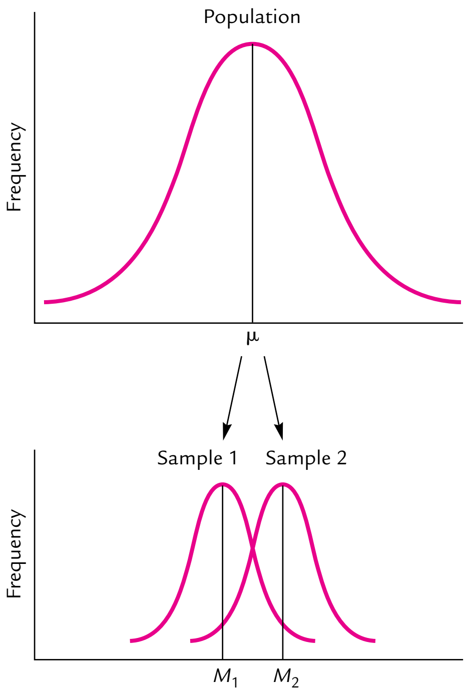


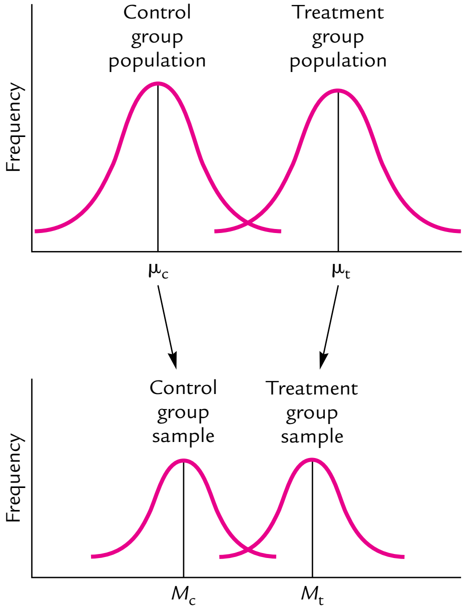
FIGURE 14-1 Line graphs showing the relationship between samples and population, assuming that the treatment had no effect on the dependent variable (M_1, mean of Sample 1; M_2, mean of Sample 2).

Consider the case in which the treatment *does* affect the scores, perhaps by shifting them upward. Figure 14-2 illustrates this situation. In the upper part of the figure is a population underlying the control group sample distribution and another one underlying the treatment group sample distribution. The population distribution underlying the treatment group is shifted upward and away from the control group population distribution. This shift could be obtained by simply adding a constant to each value in the control group distribution. This new shifted distribution resembles the old, unshifted distribution in standard deviation, but its mean is higher.

The bottom part of the figure shows two possible sample distributions—one for the control group and one for the treatment group. The scores from the control group still constitute a sample from the unshifted distribution (left-hand upper curve in Figure 14-2), but the scores from the treatment group now constitute a sample from the shifted distribution (right-hand upper curve in Figure 14-2). The two sample means provide estimates of two *different* population means. Because of sampling error, the two sample means might or might not differ even though a difference exists between the underlying population means.

Your problem (as a researcher) is that you do not know whether the treatment really had an effect on the scores. You must decide this based on your observed sample means (which may differ by a certain amount) and the sample standard deviations. From this information, you must decide whether the two sample means were drawn from the same population (the treatment had no effect on the sample scores) or from two different populations (the treatment shifted the scores relative to scores from the control group). Inferential statistics help you make this decision.

FIGURE 14-2 Line graphs showing the relationship between samples and population, assuming that the treatment had an effect on the dependent variable.



These two possibilities (different or the same populations) can be viewed as statistical hypotheses to be tested. The hypothesis that the means were drawn from the same population (i.e., $\mu_c = \mu_t$) is referred to as the *null hypothesis* (H_0). The hypothesis that the means were drawn from different populations ($\mu_c \neq \mu_t$) is called the *alternative hypothesis* (H_1).

Inferential statistics use the characteristics of the two samples to evaluate the validity of the null hypothesis. Put another way, they assess the probability that the means of the two samples would differ by the observed amount or more if they had been drawn from the same population of scores. If this probability is sufficiently small (i.e., if it is very unlikely that two samples this different would be drawn by chance from the same population), then the difference between the sample means is said to be *statistically significant*, and the null hypothesis is rejected.

When you reject the null hypothesis, you are concluding that the two samples did not come from populations having the same mean. In this example, this implies that your treatment had an effect on your dependent measure: it shifted the distribution of treatment-group scores away from that of the control group.

Statistical Errors

When making a comparison between two sample means, there are two possible states of affairs (the null hypothesis is true or it is false) and two possible decisions you can make (to reject the null hypothesis or not to reject it). In combination, these conditions lead to four possible outcomes, as shown in Table 14-1. The labels across the top of Table 14-1 indicate the two states of affairs, and those in the left-hand column indicate the two possible decisions. Each box represents a different combination of the two conditions.

TABLE 14-1 **Statistical Errors**

	TRUE STATE OF AFFAIRS	
DECISION	H_0 *True*	H_0 *False*
Reject H_0	Type I error	Correct decision
Do not reject H_0	Correct decision	Type II error

The lower left-hand box represents the situation in which the null hypothesis is true (the independent variable had no effect), and you correctly decide not to reject the null hypothesis. This is a disappointing outcome, but at least you made the right decision.

The upper left-hand box represents a more disturbing outcome. Here the null hypothesis is again true, but you have incorrectly decided to reject the null hypothesis. In other words, you decided that your independent variable had an effect when in fact it did not. In statistics this mistake is called a **Type I error**. In signal-detection experiments, the same kind of mistake is called a "false alarm" (saying that a stimulus was present when actually it was not).

The lower right-hand box in Table 14-1 represents a second kind of error. In this case, the null hypothesis is false (the independent variable did have an effect), but you have incorrectly decided not to reject the null hypothesis. This is called a **Type II error** and represents the case in which you concluded your independent variable had no effect when it really did have one. In signal-detection experiments, such an outcome is called a "miss" (not detecting a stimulus that was present).

Ideally, you would like to minimize the probability of making either a Type I or a Type II error. Unfortunately, some of the things that you can do to minimize a Type I error actually increase the probability of a Type II error, and vice versa.

Statistical Significance

If both of your samples came from the same population (or from populations having the same mean), then the null hypothesis is true, and any difference between the sample means reflects nothing more than sampling error. The actual difference between your sample means may be just such a chance difference, or it may reflect a real difference between the means of the populations from which the samples were drawn. Which of these is the case? To help you decide, you can compute an inferential statistic to determine the probability of obtaining a difference between sample means as large as or larger than the difference you actually got, under the assumption that the null hypothesis is true. If this probability is low enough, you reject the null hypothesis because you would be unlikely to have obtained the difference you did simply through sampling error.

To determine this probability, you calculate an *observed value* of your inferential statistic. This observed value is compared to a *critical value* of that statistic (normally found in a statistical table such as those in the Appendix, for example, Table 2). Ultimately, you will make your decision about rejecting the null hypothesis based on whether or not the observed value of the statistic meets or exceeds the critical value. As stated, you want to be able to reduce the probability of committing a Type I error. The probability of committing a Type I error depends on the criterion you use to accept or reject the null hypothesis. This criterion, known as the **alpha level (α)**, represents the probability that a difference at least as large as the observed difference between your sample means could have occurred purely through sampling error. The alpha level that you adopt (along with the degrees of freedom) also determines the critical value of the statistic that you are using. The smaller the value of alpha, the larger the critical value. Alpha is the probability of a Type I error. The smaller you make alpha, the less likely you are to make a Type I error. In theory, you can reduce the probability of making a Type I error to any desired level. For example, you could average less than one Type I error in 1 million experiments by choosing an alpha value of .000001. There are good reasons, discussed later, why you do not ordinarily adopt such a conservative alpha level.

By convention, the minimum acceptable alpha has been set at .05 (5 chances in 100 that sampling error alone could have produced a difference at least as large as the one observed). The particular level of alpha you adopt is called the level of *significance*. If the difference between means yields an observed value of a statistic that meets or exceeds the critical value of your inferential statistic, you declare that difference to be *statistically significant*.

The strategy of looking up the critical value of a statistic in a table and then comparing the obtained value with this critical value was developed in an era when most computations had to be done by hand, making it exceedingly difficult to find the probability with which a value equal to or larger than the obtained value of the test statistic would occur by chance when the null hypothesis is true. These days, most statistical analyses are conducted using computerized statistical packages that usually provide the exact probability value p along with the obtained value of the test statistic. You can directly compare this obtained p with your selected alpha level and avoid having to use the relevant table of critical values of your test statistic. If the obtained p is less than or equal to alpha, your comparison is statistically significant.

One-Tailed Versus Two-Tailed Tests

The critical values of a statistic depend on such factors as the number of observations per treatment, the number of treatments, and the desired alpha level. They also depend on whether the test is one-tailed or two-tailed.

Figure 14-3 shows two examples of the sampling distribution for the z statistic. This distribution is normal and therefore symmetric about the mean. The left distribution shows the **critical region** (shaded area) for a *one-tailed test*, assuming alpha has been set to .05. This region contains 5% of the total area under the curve, representing the 5% of extreme cases in this region whose *z* scores occur by chance with a probability of .05 or less. The *z* values falling into this critical region are judged to be statistically significant.

The right distribution in Figure 14-3 shows the *two* critical regions for the *two-tailed test*, using the same .05 alpha value. To keep the probability at .05, the total percentage of cases found in the two tails of the distribution must equal 5%. Thus, each critical region must contain 2.5% of the cases. Consequently, the *z* scores required to reach statistical significance must be more extreme than was the case for the one-tailed test.

You would conduct a one-tailed test if you were interested only in whether the obtained value of the statistic falls in one tail of the sampling distribution for that statistic. This is usually the case when your research hypotheses are directional. For example, you may want to know whether a new therapy is measurably *better* than the standard one. However, if the new therapy is not better, then you really do not care whether it is simply as good as the standard method or is actually worse. You would not use it in either case.

In contrast, you would conduct a two-tailed test if you wanted to know whether the new therapy was either better *or* worse than the standard method. In that case, you need to check whether your obtained statistic falls into either tail of the distribution.

The major implication of all this is that for a given alpha level you must obtain a greater difference between the means of your two treatment groups to reach statistical significance if you use a two-tailed test than if you use a one-tailed test. The one-tailed test is therefore more likely to detect a real difference if one is present (i.e., it is more powerful). However, using the one-tailed test means giving up any information about the reliability of a difference in the other, untested direction.

The use of one-tailed versus two-tailed tests has been a controversial topic among statisticians. Strictly speaking, you must choose which version you will use *before* you see the data. You must base your decision on such factors as practical considerations (as in the therapy example), your hypothesis, or previous knowledge. If you wait until after you have seen the data and then base your decision on the direction of the

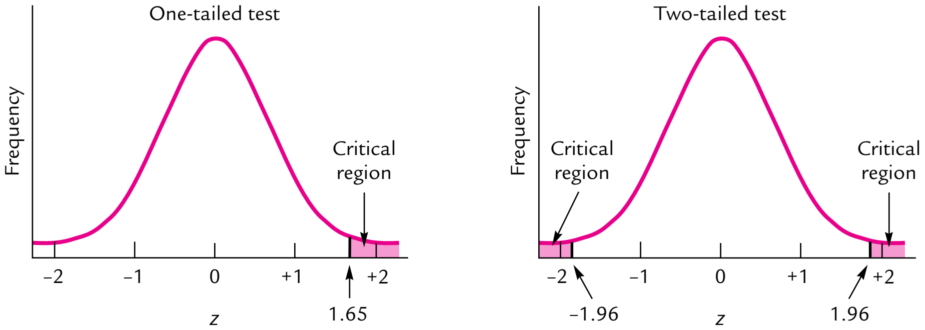


FIGURE 14-3 Graphs showing critical regions for one-tailed and two-tailed tests of statistical significance.

obtained outcome, your actual probability of falsely rejecting the null hypothesis will be greater than the stated alpha value. You have used information contained in the data to make your decision, but that information may itself be the result of chance processes and unreliable.

If you conduct a two-tailed test and then fail to obtain a statistically significant result, the temptation is to find some excuse why you "should have done" a one-tailed test. You can avoid this temptation if you adopt the following rule of thumb: Always use a two-tailed test unless there are compelling *a priori* reasons not to.

QUESTIONS TO PONDER

1. What is the general logic behind inferential statistics?
2. How are Type I and Type II errors related?
3. What does statistical significance mean?
4. When should you use a one-tailed or a two-tailed test?

PARAMETRIC STATISTICS

As previously noted, there are two types of inferential statistics: parametric and nonparametric. The type that you apply to your data depends on the scale of measurement used and how your data are distributed. This section discusses parametric inferential statistics.

Assumptions Underlying a Parametric Statistic

Three assumptions underlie parametric inferential tests (Gravetter & Wallnau, 2010): (1) The scores have been sampled randomly from the population, (2) the sampling distribution of the mean is normal, and (3) the within-groups variances are homogeneous. Assumption 3 means the variances of the different groups are highly similar. In statistical inference, the independent variable is assumed to affect the mean but not the variance.

Serious violation of one or more of these assumptions may bias the statistical test. Such bias will lead you to commit a Type I error either more or less often than the stated alpha probability and thus undermine the value of the statistic as a guide to decision making. We examine the effects of violations of these assumptions later in more detail during a discussion of the statistical technique known as the *analysis of variance*.

Inferential Statistics with Two Samples

Imagine that you have conducted a two-group experiment on whether "death-qualifying" a jury (i.e., removing any jurors who could not vote for the death penalty) affects how simulated jurors perceive a criminal defendant. Participants in your experimental group were death qualified whereas those in your control group were not.

Participants then rated on a scale from 0 to 10 the likelihood that the defendant was guilty as charged of the crime. You run your experiment and then compute a mean for each group. You find that the two means differ from one another (the experimental group mean is 7.2, and the control group mean is 4.9).

Your means may represent a single population and differ only because of sampling error. Or your means may reliably represent two different populations. Your task is to determine which of these two conditions is true. Is the observed difference between means reliable, or does it merely reflect sampling error? This question can be answered by applying the appropriate statistical test, which in this case is a *t* test.

The *t* Test

Use the ***t* test** when your experiment includes only two levels of the independent variable (as in the jury example). Special versions of the *t* test exist for designs involving independent samples (e.g., randomized groups) and for those involving correlated samples (e.g., matched-pairs designs and within-subjects designs).

The t Test for Independent Samples You use the ***t* test for independent samples** when you have data from two groups of participants who were assigned at random to the two groups. The test comes in two versions, depending on the error term selected. The *unpooled* version computes an error term based on the standard error of the mean provided separately by each sample. The *pooled* version computes an error term based on the two samples combined, under the assumption that both samples come from populations having the same variance. The pooled version may be more sensitive to any effect of the independent variable, but it should be avoided if there are large differences in sample sizes and standard errors. Under these conditions, the probability estimates provided by the pooled version may be misleading.

The t Test for Correlated Samples When the two means being compared come from samples that are not independent of one another, the formula for the *t* test must be adjusted to take into account any correlation between scores; the adjusted version is called the ***t* test for correlated samples**. In such cases, the scores from the two samples come in pairs arising from two observations of the same variable on the same participant or from single observations taken on each of a matched pair of participants. Within-subjects and matched-pairs experimental designs and some correlational designs meet this requirement.

The *t* test for correlated samples produces a larger *t* value than the *t* test for independent samples when applied to the same data *if* the scores from the two samples are at least moderately correlated, and this tends to make the correlated samples test more sensitive to any effect of the independent variable. However, this advantage tends to be offset by the correlated sample *t* test's smaller degrees of freedom [equal to $n - 1$, where n is the number of *pairs* of scores, as opposed to the $(n_1 - 1) + (n_2 - 1)$ degrees of freedom of the *t* test for independent samples, where n_1 and n_2 are the number of scores in the two samples]. When the correlation between samples is 0, the *t* values given by the correlated samples and independent samples *t* tests (pooled version) are identical; with its reduced degrees of freedom, the correlated samples *t* test will then be less able than the independent samples *t* test to detect any effect of the independent variable.

An Example from the Literature: Contrasting Two Groups

Spinal cord injuries (SCI) represent a major source of physical disabilities (Hess, Marwitz, & Kreutzer, 2003). SCIs are often the result of automobile accidents or falls that involve rapid deceleration of the body and may result in mild traumatic brain injury (MTBI). Hess et al. note that when a patient with an SCI is rushed into the emergency room, the possibility that MTBI exists is often overlooked because of the seriousness of SCIs. Often, patients with SCI show cognitive impairments normally associated with MTBI, such as memory loss, attention deficits, and problems with processing information (Hess et al., 2003). The problem is that it is sometimes difficult to determine whether cognitive impairments are the result of MTBI or the emotional trauma associated with SCI.

David Hess, Jennifer Marwitz, and Jeffrey Kreutzer (2003) conducted a study to differentiate between patients with MTBI (without SCI) and patients with SCI. Participants were patients with SCI or MTBI who had been treated at a medical center. Participants' neuropsychological functioning was measured using a battery of tests assessing attention (two tests), motor speed, verbal learning, verbal memory (two tests), visuospatial skills, and word fluency. Mean scores were computed on each measure for patients with SCI and MTBI. Hess et al. used a series of *t* tests to determine if the SCI and MTBI patients differed significantly on any of the neuropsychological tests. They found significant differences between the two groups on 5 of the 10 tests. The results (shown in Table 14-2) showed that, as a rule, patients with SCI performed better than patients with MTBI. They also found that a high percentage of SCI patients showed significant impairment on several of the cognitive measures (even though they scored better than the MTBI patients). Hess et al. suggest that SCI patients might benefit from a comprehensive rehabilitation program that targets cognitive functioning as well as emotional well-being.

As presented, the data in Table 14-2 do not make much sense. All that you have are means and a *t* value (with its degrees of freedom) for each measure. You must decide if the *t* values are large enough to warrant a conclusion that the observed differences are statistically significant.

After calculating a *t* score, you compare its value with a critical value of *t* found in Table 2 of the Appendix. Before you can evaluate your obtained *t* value, however, you must obtain the degrees of freedom (for the between-subjects *t* test, $df = N - 2$, where *N* is the total number of subjects in the experiment).

Once you have obtained the degrees of freedom (these are shown in parentheses in the fourth column of Table 14-2), you compare the obtained *t* score with the tabled critical value, a process requiring two steps. In Table 2 of the Appendix, first read down the column labeled "Degrees of Freedom" and find the number matching your degrees of freedom. Next, find the column corresponding to the desired alpha level (labeled "Alpha Level"). The critical value of *t* is found at

TABLE 14-2 **Means and *t* Values From the Five Significant Differences Found by Hess et al. (2003)**

TEST	SCI	MTBI	*t(df)*
Written attention test	41.6	30.4	2.40 (18)
Motor speed	91.4	126.1	−2.20 (31)
Verbal learning	47.1	37.9	2.40 (34)
Verbal memory (immediate recall)	25.9	18.7	3.16 (49)
Verbal memory (delayed recall)	21.4	10.7	4.73 (44)

the intersection of the degrees of freedom (row) and alpha level (column) of your test. If your obtained *t* score is equal to or greater than the tabled *t* score, then the difference between your sample means is statistically significant at the selected alpha level.

In some instances, you may find that the table you have does not include the degrees of freedom that you have calculated (e.g., 44). If this occurs, you can use the next *lower* degrees of freedom in the table. With 44 degrees of freedom, you would use the entry for 40 degrees of freedom in the table.

If you are conducting your *t* tests on a computer, most statistical packages will compute the exact *p* value for the test, given the obtained *t* and degrees of freedom. In that case, simply compare your obtained *p* values to your chosen alpha level. If *p* is less than or equal to alpha, the difference between your groups is statistically significant at the stated alpha level.

The *z* Test for the Difference Between Two Proportions

In some research, you may have to determine whether two proportions are significantly different. In a jury simulation in which participants return verdicts of guilty or not guilty, for example, your dependent variable might be expressed as the proportion of participants who voted guilty. A relatively easy way to analyze data of this type is to use a ***z* test for the difference between two proportions**. The logic behind this test is essentially the same as for the *t* tests. The difference between the two proportions is evaluated against an estimate of error variance.

QUESTIONS TO PONDER

1. What are the assumptions underlying parametric statistics?
2. Which parametric statistics would you use to analyze data from an experiment with two independent groups?
3. Which parametric statistic is appropriate for a matched two-group design?

Beyond Two Groups: Analysis of Variance (ANOVA)

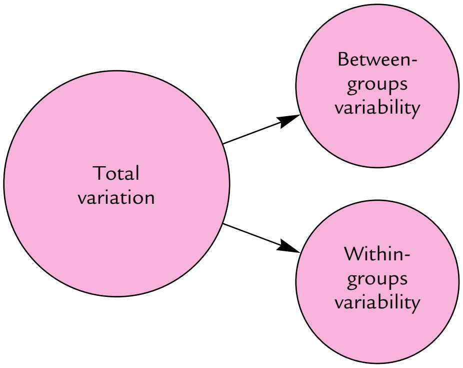
When your experiment includes more than two groups, the statistical test of choice is **analysis of variance (ANOVA)**. As the name implies, ANOVA is based on the concept of analyzing the variance that appears in the data. For this analysis, the variation in scores is divided, or *partitioned*, according to the factors assumed to be responsible for producing that variation. These factors are referred to as *sources of variance*. The next sections describe how variation is partitioned into sources and how the resulting source variations are used to calculate a statistic called the *F* ratio. The *F* ratio is ultimately checked to determine whether the variation among means is statistically significant.

Partitioning Variation The value of any particular score obtained in a between-subjects experiment is determined by three factors: (1) characteristics of the subject at the time the score was measured, (2) measurement or recording errors (together called *experimental error*), and (3) the value of the independent variable (assuming the independent variable is effective). Because subjects differ from one another (Factor 1) and because measurement error fluctuates (Factor 2), scores will vary from one another even when all subjects are exposed to the same treatment conditions. Scores will vary even more if subjects are exposed to different treatment conditions and the independent variable is effective.

Figure 14-4 shows how the total variation in the scores from a given experiment can be partitioned into two sources of variability (between-groups variability and within-groups variability). Notice that the example begins with a total amount of variability among scores. Again, this total amount of variability may be attributable to one or more of three factors: your independent variable, individual differences, and experimental error (Gravetter & Wallnau, 2010).

The first component resulting from the partition is the *between-groups variability*. The between-groups variability may be caused by the variation in your independent variable, by individual differences among the different subjects in your groups, by experimental error, or by a combination of these (Gravetter & Wallnau, 2010). The second component, the *within-groups variability*, may be attributed to error. This error

FIGURE 14-4 Partitioning total variation into between-groups and within-groups sources.



can arise from either or both of two sources: individual differences between subjects treated alike within groups and experimental error (Gravetter & Wallnau, 2010). Take note that variability caused by your treatment effects is unique to the between-groups variability.

The F Ratio The statistic used in ANOVA to determine statistical significance is the ***F* ratio**. The *F* ratio is simply the ratio of between-groups variability to within-groups variability. Both types of variability that constitute the ratio are expressed as variances. (Chapter 13 described the variance as a measure of spread.) However, statisticians perversely insist on calling the variance the *mean square*, perhaps because the term is more descriptive. Just as with the *t* statistic, once you have obtained your *F* ratio, you compare it against a table of critical values to determine whether your results are statistically significant.

The One-Factor Between-Subjects ANOVA

Use the one-factor between-subjects ANOVA when your experiment includes only one factor (with two or more levels) and has different subjects in each experimental condition. As an example, imagine you have conducted an experiment on how well participants can detect a signal against a background of noise, measured in decibels (db). Participants were exposed to different levels of background noise (no noise, 20 db, or 40 db) and asked to indicate whether or not they heard a tone. The number of times that the participant correctly stated that a tone was present represents your dependent variable. You found that participants in the no-noise group detected more of the tones (36.4) than participants in either the 20-db (23.8) or 40-db (16.0) group. Table 14-3 shows the distributions for the three groups.

Submitting your data to a one-factor between-subjects ANOVA, you obtain an *F* ratio of 48.91. This *F* ratio is now compared with the appropriate critical value of *F* in Tables 3A and 3B in the Appendix. To find the critical value, you need to use the

TABLE 14-3 Data from Hypothetical Signal-Detection Study

	NO NOISE	20 DECIBELS	40 DECIBELS
	33	22	17
	39	24	14
	41	25	19
	32	21	11
	37	27	19
$\sum X$	182	119	80
$\sum X^2$	6,684	2,855	1,328
M	36.4	23.8	16.0

degrees of freedom for both the numerator ($k - 1$, where k is the number of groups) *and* the denominator [$k(n - 1)$, where n is the number of subjects in each group] of your F ratio. In this case, the degrees of freedom for the numerator and denominator are 2 and 12, respectively.

To identify the appropriate critical value for F (at $\alpha = .05$), first locate the appropriate degrees of freedom for the numerator across the top of Table 3A. Then read down the left-hand column to find the degrees of freedom for the denominator. In this example, the critical value for F (2, 12) at $\alpha = .05$ is 3.89. Because your obtained F ratio is greater than the tabled value, you have an effect significant at $p < 05$. In fact, if you look at the critical value for F (2, 12) at α .01 (found in Table 3B), you will find that your obtained F ratio is also significant at $p < .01$.

As noted earlier, when you report a significant effect, typically you express it in terms of a ***p* value**. *Alpha* refers to the cutoff point that you adopt. In contrast, the p value refers to the actual probability of making a Type I error given that the null hypothesis is true. Hence, for this example, you would report that your finding was significant at $p < .05$ or $p < .01$. The discussion in the following sections assumes the "$p <$" notation.

Sometimes the table of the critical values of F does not list the exact degrees of freedom for your denominator. If this happens, you can approximate the critical value of F by choosing the next lower degrees of freedom for the denominator in the table. Choosing this lower value provides a more conservative test of your F ratio.

Interpreting Your F Ratio A significant F ratio tells you that at least some of the differences among your means are probably not caused by chance but rather by variation in your independent variable. The only problem, at this point, is that the F ratio fails to tell you where among the possible comparisons the reliable differences actually occur. To isolate which means differ significantly, you must conduct specific comparisons between pairs of means. These comparisons can be either planned or unplanned.

Planned Comparisons **Planned comparisons** (also known as *a priori comparisons*) are used when you have specific preexperimental hypotheses. For example, you may have hypothesized that the no-noise group would differ from the 40-db group but not from the 20-db group. In this case, you would compare the no-noise and 40-db groups and then the no-noise and 20-db groups. These comparisons are made using information from your overall ANOVA (see Keppel, 1982). Separate F ratios (each having 1 degree of freedom) or t tests are computed for each pair of means. The resulting F ratios or obtained t values are then compared to their respective critical values.

You can conduct as many of these planned comparisons as necessary. However, a limited number of such comparisons yield unique information. For example, if you found that the no-noise and 20-db groups did not differ significantly and that the 40- and 20-db groups did, you have no reason to compare the no-noise and 40-db groups. You can logically infer that the no-noise and 40-db groups differ significantly. Those comparisons that yield new information are known as *orthogonal comparisons*. Any set of means has ($k - 1$) orthogonal comparisons, where k is the number of treatments.

You can use planned comparisons in lieu of an overall ANOVA if you have highly specific preexperimental hypotheses. In this case, you would not have the information required to use the given formula for planned comparisons. A simple alternative is to conduct multiple *t* tests. You should not perform too many of these comparisons even if the relationships were predicted before you conducted your experiment. Performing multiple tests on the same data increases the probability of making a Type I error across comparisons through a process called *probability pyramiding* (discussed in the next section).

Unplanned Comparisons If you do not have a specific preexperimental hypothesis concerning your results, you must conduct **unplanned comparisons** (also known as *post hoc comparisons*). Unplanned comparisons are often "fishing expeditions" in which you are simply looking for any differences that might emerge. In experiments with many levels of an independent variable, you may be required to perform a fairly large number of unplanned comparisons to fully analyze the data.

Two types of error must be considered when making many comparisons: **per-comparison error** and **familywise error**. Per-comparison error is the alpha for each comparison between means. If you set an alpha level of .05, the per-comparison error rate is .05. The familywise error rate (Keppel, 1982) takes into account the increasing probability of making at least one Type I error as the number of comparisons increases (i.e., probability pyramiding). You compute familywise error with the following formula:

$$\alpha_{FW} = 1 - (1 - \alpha)^c$$

where *c* is the number of comparisons made and is your per-comparison error rate. For example, if you are making four comparisons ($c = 4$) and $\alpha = .05$, then $\alpha_{FW} = 1 - (1 - .05)^4 = 1 - .95^4 = 1 - .815 = .185$. Thus the chance of getting at least one significant difference by chance in four comparisons is more than three times the stated alpha level, assuming that only chance is at work to produce these differences.

Special tests can be applied to control familywise error, but it is beyond the scope of this chapter to discuss each of them individually. Table 14-4 lists the tests most often used to control familywise error and gives a brief description of each. For more information about these tests, see Keppel (1982, Chapter 8).

Sample Size You can still use an ANOVA if your groups contain unequal numbers of subjects, but you must use adjusted computational formulas. The adjustments can take one of two forms, depending on the reasons for unequal within-cell sample sizes.

Unequal sample sizes may simply be a by-product of the way that you conducted your experiment. If you conducted your experiment by randomly distributing your materials to a large group, for example, you would not be able to keep the sample sizes equal. In such cases, unequal sample sizes do not result from the properties of your treatment conditions.

Unequal sample sizes also may result from the effects of your treatments. If one of your treatments is painful or stressful, participants may drop out of your experiment

TABLE 14-4 Post Hoc Tests

TEST	USE	COMMENTS[a]
Scheffé test	To keep familywise error rate constant regardless of the number of comparisons to be made	Very conservative test; Scheffé correction factor corrects for all possible comparisons, even if not all are made
Dunnett test	To contrast several experimental groups with a single control group	Not as conservative as the Scheffé test because only the number of comparisons made is considered in the familywise error rate correction
Tukey-a HSD test	To hold the familywise error rate constant over an entire set of two-group comparisons	Not as conservative as the Scheffé test for comparisons between pairs of means; less powerful than the Scheffé for more complex comparisons
Tukey-b WSD test	Alternative Tukey test	Not as conservative as Tukey's HSD test, but more conservative than the Newman–Keuls test
Newman–Keuls test	To compare all possible pairs of means and control per-comparison error rate	Less conservative than the Tukey test; critical value varies according to the number of comparisons made
Ryan's Test (REGWQ)	Modified Newman–Keuls test in which critical values decrease as the range between the highest and lowest means decreases	Controls familywise error better than the Newman–Keuls test but is less powerful than the Newman–Keuls test
Duncan test	To compare all possible pairs of means	Computed in the same way as the Newman–Keuls test; with more than two means to be compared, it is less conservative than the Newman–Keuls
Fisher test	To compare all possible combinations of means	Powerful test that does not overcompensate to control familywise error rate; no special correction factor used; significant overall *F* ratio justifies comparisons

[a] A conservative test is one with which it is more difficult to achieve statistical significance than with a less conservative test. "Power" refers to the ability of a test to reject the null hypothesis when the null hypothesis is false.

SOURCE: Information in this table was summarized from Keppel, 1982, pp. 153–159; Pagano, 2010; Winer, 1971; and information found at http://www2.chass.ncsu.edu/garson/pa765/anova.htm.

because of the aversive nature of that treatment. Death of animals in a group receiving highly stressful conditions is another example of subject loss related to the experimental manipulations that result in unequal sample sizes.

Unweighted Means Analysis If you end up with unequal sample sizes for reasons *not* related to the effects of your treatments, one solution is to equalize the groups by randomly discarding the excess data from the larger groups. Even then, discarding data may not be a good idea, especially if the sample sizes are small to begin with. The loss of data inevitably reduces the power of your statistical tests.

Rather than dropping data, you could use an *unweighted means analysis* that involves a minor correction to the ANOVA. This analysis gives each group in your design equal weight in the analysis, despite unequal group sizes.

Weighted Means Analysis If the inequality in sample sizes was planned or reflects actual differences in the population, you should use a *weighted means analysis* (Keppel, 1973). In a weighted means analysis, each group mean is weighted according to the number of subjects in the group. As a result, means with higher weightings (those from larger groups) contribute more to the analysis than do means with lower weights. See Keppel (1973, 1982) or Gravetter and Wallnau (2010) for more information about unequal sample size in ANOVA.

The One-Factor Within-Subjects ANOVA

If you used a multilevel within-subjects design in your experiment, the statistical test to use is the *one-factor within-subjects* ANOVA. As in a between-subjects analysis, the between-treatments sum of squares can be affected by the level of the independent variable and by experimental error (Gravetter & Wallnau, 2010). However, unlike the between-subjects case, individual differences no longer contribute to the between-treatments sum of squares because the same subjects are in each experimental treatment group. The within-subjects source of variance(s) also can be partitioned into two factors: variability within a particular treatment (i.e., different subjects reacting differently to the same treatment) and experimental error.

The contribution of individual differences is estimated by treating subjects as a factor in the analysis (S). You then subtract S from the usual within-groups variance. This subtraction reduces the amount of error in the denominator of the F ratio, thus making the F ratio more sensitive to the effects of the independent variable—a major advantage.

The Latin Square ANOVA Latin square designs are used to counterbalance the order in which subjects receive treatments in within-subjects experiments (see Chapter 10). The carryover effects contained in the Latin square design tend to inflate the error term used to calculate your F ratio. Consequently, they must be removed before you calculate F. This is done by treating practice effects as a factor in the analysis and removing their effects from the error term. For more information on the Latin square ANOVA, see Keppel (1982, pp. 385–391).

Interpreting Your F Ratio A significant overall F ratio tells you that significant differences exist among your means, but, as usual, it does not tell you where these significant differences occur. To determine which means differ, you must further analyze your data. The tests used to compare your means are similar to those used in the between-subjects analysis. Once again, they can be either planned or unplanned.

QUESTIONS TO PONDER

1. When would you need to use a one-factor ANOVA rather than a t test to analyze your data?
2. Why should you normally use ANOVA to analyze data from more than two treatments, rather than conducting multiple t tests?
3. When would you do a planned versus an unplanned comparison, and why?
4. What is the difference between weighted and unweighted means analysis, and when would you use each?
5. What is a post hoc test and what does it control?
6. What are Latin square designs? What are they used for?

The Two-Factor Between-Subjects ANOVA

Chapter 10 discussed the two-factor between-subjects design. In this design, you include two independent variables and randomly assign different subjects to each condition. In addition, you combine independent variables across groups so that you can extract the independent effect of each factor (the main effects) and the combined effect of the two factors (interaction) on the dependent variable. (If you are unclear about the meanings of these terms, review Chapter 10.) The analysis appropriate to data from this design is the *two-factor between-subjects ANOVA*. This ANOVA is necessarily more complicated than a one-factor ANOVA because it must determine the statistical significance of each main effect and of the interaction as well.

Main Effects and Interactions If you find significant main effects and a significant interaction in your experiment, you must be careful about interpreting the main effects. When you interpret a main effect, you are suggesting that your independent variable has an effect on the dependent variable, regardless of the level of your other independent variable. The presence of an interaction provides evidence to the contrary. The interaction shows that neither of your independent variables has a simple, independent effect. Consequently, you should avoid interpreting main effects when an interaction is present.

You should also be aware that certain kinds of interactions can cancel out the main effects. The independent variables may have been effective, and yet the statistical analysis will fail to reveal statistically significant main effects for these factors. To see how this can happen, imagine you have conducted a two-factor experiment

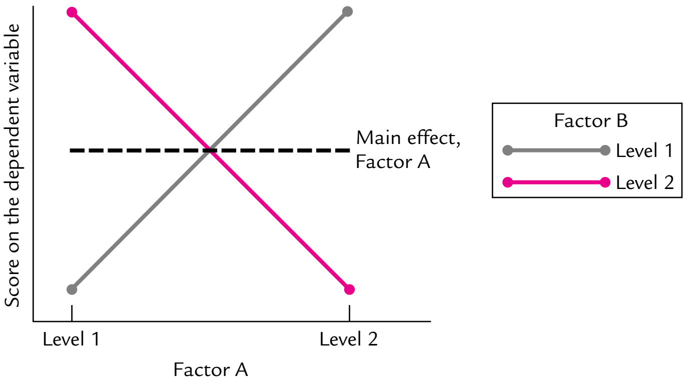


FIGURE 14-5 Graph showing a two-way interaction that masks main effects.

with two levels of each factor. Figure 14-5 graphs the cell means for this hypothetical experiment.

The diagonal lines depict the functional relationship between Factor A and the dependent variable at the two levels of Factor B. The fact that the lines form an X (rather than being parallel) indicates the presence of an interaction. Notice that Factor A strongly affects the level of the dependent variable at *both* levels of Factor B but that these effects run in opposite directions.

The dashed line in Figure 14-5 represents the main effect of Factor A, computed by averaging the upper and lower points to collapse across the levels of Factor B. This dashed line is horizontal, indicating that there is no change in the dependent variable across the two levels of Factor A (collapsed over Factor B). Although Factor A has strong effects on the dependent variable at each level of Factor B, its average (main) effect is zero.

Logically, if the interaction of two variables is significant, then the two variables themselves have reliable effects. Consequently, if you have a significant interaction, do not interpret the main effects. The effects of the factors involved in the interaction are reliable whether or not the main effects are statistically significant.

Finally, most of the time you are more interested in the significant interaction than in main effects, even before your experiment is conducted. Hypothesized relationships among variables are often stated in terms of interactions. Interactions tend to be inherently more interesting than main effects. They show how changes in one variable alter the effects on behavior of other variables.

Sample Size Just as with a one-factor ANOVA, you can compute a multifactor ANOVA with unequal sample sizes. The unweighted means analysis can be conducted on a design with two or more factors (the logic is the same). For details on modifications to the basic two-factor ANOVA formulas for weighted means and unweighted means analyses, see Keppel (1973, 1982).

ANOVA for a Two-Factor Between-Subjects Design: An Example An experiment conducted by Doris Chang and Stanley Sue (2003) provides an excellent example of the application of ANOVA to the analysis of data from a two-factor experiment.

Chang and Sue were interested in investigating how the race of a student affected a teacher's assessments of the student's behavior and whether those assessments were specific to certain types of issues. Teachers (163 women and 34 men) completed a survey on which they were asked to evaluate the behavior of three hypothetical children. Each survey included a photograph of an Asian-American, an African-American, or a Caucasian child. The survey also included a short description of the child's behavior, which was depicted as falling into one of three "problem" types: (1) "overcontrolled" (anxious to please and afraid of making mistakes), (2) "undercontrolled" (disobedient, disruptive, and easily frustrated), or (3) "normal" (generally follows rules, fidgets only occasionally, etc.). These two variables comprise the two independent variables in a 3 (race of child) × 3 (problem type) factorial design. The survey also included several measures on which teachers evaluated the child's behavior (e.g., seriousness, how typical the behavior was, attributions for the causes of the behavior, and academic performance).

We limit our discussion of the results to one of the dependent variables: typicality of the behavior. The data were analyzed with a two-factor ANOVA. The results showed a significant main effect of problem type, F (2, 368) = 46.19, $p < .0001$. Normal behavior (M = 6.10) was seen as more typical than either undercontrolled (M = 4.08) or overcontrolled (M = 4.34) behavior. The ANOVA also showed a statistically significant race by problem–type interaction, F (4, 368) = 7.37, $p < .0001$.

Interpreting the Results This example shows how to interpret the results from a two-factor ANOVA. First, consider the two main effects. There was a significant effect of problem type on typicality ratings. Normal behavior was rated as more typical than overcontrolled or undercontrolled behavior. If this were the only significant effect, you could then conclude that race of the child had no effect on typicality ratings because the main effect of race was not statistically significant. However, this conclusion is not warranted because of the presence of a significant interaction between race of learner and problem type.

The presence of a significant interaction suggests that the relationship between the two independent variables and your dependent variable is complex. Figure 14-6 shows the data contributing to the significant interaction in the Chang and Sue (2003) experiment. Analyzing a significant interaction like this one involves making comparisons among the means involved.

Because Chang and Sue (2003) predicted the interaction, they used planned comparisons (t tests) to contrast the relevant means. The results showed that the typicality of the Asian-American child's behavior was evaluated very differently from that of the Caucasian child and African-American child. Teachers saw the normal behavior of the Asian-American child as less typical than the normal behavior of either the Caucasian or African-American child. Teachers saw the overcontrolled behavior by the Asian-American child as more typical than the same behavior attributed to the African-American or Caucasian child. The undercontrolled behavior was seen as less typical for the Asian-American child than for the African-American and Caucasian children, respectively. So the race of the

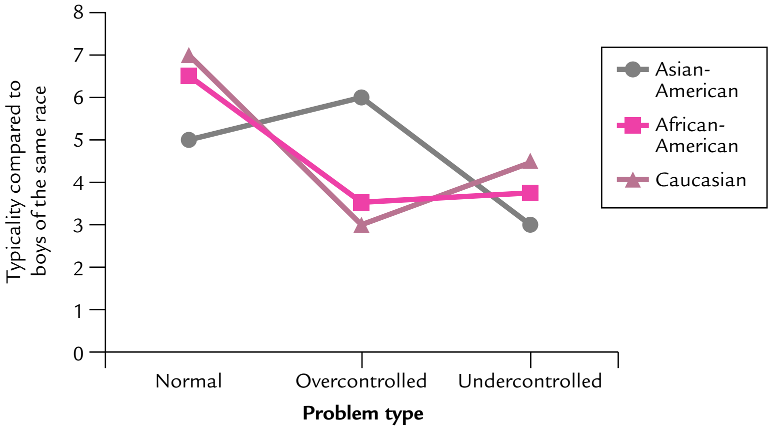


FIGURE 14-6 Graph showing an interaction between race and problem type.
SOURCE: Chang and Sue, 2003; reprinted with permission.

child did affect how participants rated the typicality of a behavior, but the nature of that effect depended on the type of behavior attributed to the child.

The Two-Factor Within-Subjects ANOVA

All subjects in a within-subjects design with two factors are exposed to every possible combination of levels of your two independent variables. Use a *two-factor within-subjects ANOVA* to analyze these designs. This analysis applies the same logic developed for the one-factor within-subjects ANOVA. As in the one-factor case, subjects are treated as a factor along with your manipulated independent variables.

The major difference between the one- and two-factor within-subjects ANOVA is that you must consider the interaction between each of your independent variables and the subjects factor (A × S and B × S), in addition to the interaction between your independent variables (A × B). Because the basic logic and interpretation of results from a within-subjects ANOVA are essentially the same as for the between-subjects ANOVA, a complete example is not given here. A complete example of the two-factor within-subjects ANOVA can be found in Keppel (1973).

Mixed Designs

In some situations, your research may call for a design mixing between-subjects and within-subjects components. This design was discussed briefly in Chapter 11. If you use such a design (known as a *mixed* or *split-plot design*), you can analyze your data with an ANOVA. The computations involve calculating sums of squares for the between factor and for the within factor.

The most complex part of the analysis is the selection of an error term to calculate the *F* ratios. The within-groups mean square is used to calculate the between-subjects *F* whereas the interaction of the within factor with the within-groups variance is used to evaluate both the within-subjects factor and the interaction between the within-subjects and between-subjects factors. Keppel (1973, 1982) provides an excellent discussion of this analysis and a complete worked example.

Higher-Order and Special-Case ANOVAs

Variations of ANOVA exist for just about any design used in research. For example, you can include three or four factors in a single experiment and analyze the data with a *higher-order* ANOVA. In a three-factor ANOVA, for example, you can test three main effects (A, B, and C), three two-way interactions (AB, AC, and BC), and a three-way interaction (ABC). As you add factors, however, the computations become more complex and probably should not be done by hand. In addition, as discussed in Chapter 10, it may be difficult to interpret the higher-order interactions with more than four factors.

A special ANOVA is used when you have included a continuous correlational variable in your experiment (such as age). This type of ANOVA, called the **analysis of covariance (ANCOVA)**, allows you to examine the relationship between experimentally manipulated variables while controlling another variable that may be correlated with them. Keppel (1973, 1982) provides clear discussions of these analyses and other issues relating to ANCOVA.

ANOVA: Summing Up

To summarize, ANOVA is a powerful parametric statistic used to analyze one-factor experiments (either within-subjects or between-subjects) with more than two treatments and to analyze multifactor experiments. It is intended for use when your dependent variable is scaled on at least an interval scale. The assumptions that apply to the use of parametric statistics in general (such as homogeneity of variance and normally distributed sampling distribution) apply to ANOVA.

ANOVA involves forming a ratio between the variance (mean square) caused by your independent variable plus experimental error and the variance (mean square) caused by experimental error alone. The resulting score is called an *F* ratio. A significant *F* ratio tells you that at least one of your means differs from the other means. Once a significant effect is found, you then perform more detailed analyses of the means contributing to the significant effect in order to determine where the significant differences occur. These tests become more complicated as the design of your experiment becomes more complex.

QUESTIONS TO PONDER

1. If you have two independent variables in your experiment, what type of ANOVA should be used to analyze your data?
2. What are main effects and interactions, and how are they analyzed?
3. What is a higher-order ANOVA? What difficulties arise as the number of orders increases?
4. What is ANCOVA, and what does it do that ANOVA does not do?

NONPARAMETRIC STATISTICS

Thus far, this discussion has centered on parametric statistical tests. In some situations, however, you may not be able to use a parametric test. When your data do not meet the assumptions of a parametric test or when your dependent variable was scaled on a nominal or ordinal scale, consider a nonparametric test. This section discusses three nonparametric tests: chi-square, the Mann–Whitney *U* test, and the Wilcoxon signed ranks test. You might consider using many other nonparametric tests. For a complete description of these, see Siegel and Castellan (1988). Table 14-5 summarizes some information on these and other nonparametric tests.

Chi-Square

When your dependent variable is a dichotomous decision (such as yes/no or guilty/not guilty) or a frequency count (such as how many people voted for Candidate A and how many for Candidate B), the statistic of choice is **chi-square (χ^2)**. Versions of chi-square exist for studies with one and two variables. This discussion is limited to the two-variable case. For further information on the one-variable analysis, see either Siegel and Castellan (1988) or Roscoe (1975).

Chi-Square for Contingency Tables *Chi-square for contingency tables* (also called the *chi-square test for independence*) is designed for frequency data in which the relationship, or contingency, between two variables is to be determined. In a voter preference study, for example, you might have measured sex of respondent in addition to candidate preference. You may want to know whether the two variables are related or independent. The chi-square test for contingency tables compares your *observed cell frequencies* (those you obtained in your study) with the *expected cell frequencies* (those you would expect to find if chance alone were operating).

A study reported by Herbert Harari, Oren Harari, and Robert White (1985) provides an excellent example of the application of the chi-square test to the analysis of frequency data. Harari et al. investigated whether male participants would help the victim of a simulated rape. Previous research on helping behavior suggested that individuals are less likely to help someone in distress if they are with others than if they are alone. Harari et al. conducted a field investigation of this effect. Participants (either walking alone or in noninteracting groups) were exposed to a mock rape (a male confederate of the experimenters grabs a female confederate and drags her into some bushes). Observers recorded whether participants helped the female rape victim. Table 14-6 shows the frequencies of participants helping under the two conditions. The results from a chi-square test performed on these data showed a significant relationship between the decision to offer help and whether participants were alone or in groups. Participants in groups were actually more likely to help than those who were alone.

Limitations of Chi-Square A problem arises if any of your expected cell frequencies is less than five. In such cases, the value of chi-square may be artificially inflated (Gravetter & Wallnau, 2010). You have three options to deal with this problem. First, you could include more subjects to increase your sample size. Second, you could

TABLE 14-5 **Nonparametric Tests**

TEST	MINIMUM SCALE OF MEASUREMENT	COMMENTS
One-Sample Tests		
Binomial	Nominal	
Chi-square	Nominal	
Kolmogorov–Smirnov	Ordinal	Can be used as a more powerful alternative to chi-square
Two Independent Samples		
Chi-square	Nominal	
Fisher exact probability	Nominal	Alternative to chi-square when expected frequencies are small
Kolmogorov–Smirnov	Ordinal	More powerful than the Mann–Whitney *U* test
Wald–Wolfowitz runs	Ordinal	
Moses test of extreme reactions	Ordinal	Less powerful than Mann–Whitney *U* test
Randomization test	Interval	Tests the difference between means without assuming normality of data or homogeneity of variance
Mann–Whitney *U*	Ordinal or above	Good alternative to *t* test when assumptions violated
Two Related Samples		
McNemar	Nominal	Good test when you have a before–after hypothesis
Sign	Ordinal	Good when quantitative measures are not possible, but you can rank data
Wilcoxon matched pairs	Ordinal	Good alternative to *t* test when normality assumption is violated
Walsh test	Interval	Good nonparametric alternative to the *t* test; data must be distributed symmetrically
Randomization test for matched pairs	Interval	

TABLE 14-5 Nonparametric Tests *continued*

More Than Two Related Samples		
Cochran Q test	Nominal	Most useful when data fall into natural dichotomous categories
Friedman two-way ANOVA	Ordinal	
More Than Two Independent Samples		
Chi-square	Nominal	
Kruskal–Wallis one-way ANOVA	Ordinal	Good alternative to a one-factor ANOVA when assumptions are violated

SOURCE: Data from Roscoe, 1975; and Siegel and Castellan, 1988.

TABLE 14-6 Number of Participants Helping Mock Rape Victim, in Two Conditions

	HELPED	DID NOT HELP	
PARTICIPANTS IN GROUPS	34	6	40
PARTICIPANTS ALONE	26	14	40
	60	20	

SOURCE: Data from Harari, Harari and White, 1985.

combine cells (if it is logical to do so); for example, you could categorize subjects into three categories rather than five. Third, you could consider a different test. The Fisher exact probability test (see Roscoe, 1975, or Siegel & Castellan, 1988) is an alternative to chi-square when you have small expected frequencies and a 2 × 2 contingency table (Roscoe, 1975).

A significant chi-square tells you that your two variables are significantly related. In the previous example, all you know is that group size and helping are related. As with ANOVA, however, chi-square does not tell you where the significant differences occur when more than two categories of each variable exist. To determine the locus of the significant effects, you can conduct separate chi-square tests on specific cells of the contingency table.

The Mann–Whitney *U* Test

Another powerful nonparametric test is the **Mann–Whitney *U* test**. Use the Mann–Whitney *U* test when your dependent variable is scaled on at least an ordinal scale. It is also a good alternative to the *t* test when your data do not meet the assumptions of the *t* test (such as when the scores are not normally distributed or when the variances are heterogeneous).

Calculation of the Mann–Whitney *U* test is fairly simple. The first step is to combine the data from your two groups. Rank the scores (from highest to lowest) and label them according to the group to which they belong. If there is a difference between your groups, then the ranks for the scores in one group should be consistently above the ranks from the other group, rather than being randomly distributed. You calculate a *U* score for each group in your experiment, then evaluate the *lower* of the two *U* scores obtained against the critical value of *U*. If the lower of the two *U* scores is *smaller* than the tabled *U* value, you then conclude your two groups differ significantly.

The Wilcoxon Signed Ranks Test

If you conducted a single-factor experiment using a correlated-samples (related) or matched-pairs design, the **Wilcoxon signed ranks test** would be a good statistic with which to analyze your data. For this test, a difference score is calculated for each pair of scores for each subject. The resulting difference scores are then ranked (disregarding the sign of the difference score) from smallest to largest. Next, each rank is assigned a positive or negative sign, depending on whether the difference score was positive or negative. The positive and negative ranks are then summed. If the null hypothesis is true, then the two sums should be equal or very close to being equal. However, if the sums of the positive and negative ranks are very different, then the null hypothesis can be rejected. For more information on the Wilcoxon signed ranks test, see Siegel and Castellan (1988).

Parametric Versus Nonparametric Statistics

Nonparametric statistics are useful when your data do not meet the assumptions of parametric statistics. If you have a choice, choose a parametric statistic over a nonparametric one because parametric statistics are generally more powerful. That is, the parametric statistic usually provides a more sensitive test of the null hypothesis than does an equivalent nonparametric statistic.

A second problem with nonparametric statistics is that appropriate versions are not always available for complex designs. Consequently, when designing your study, you should try to scale your dependent measures so an ANOVA or other suitable parametric statistic can be used.

QUESTIONS TO PONDER

1. What is a nonparametric statistic? Under what conditions would you use one?
2. When would you use the chi-square test for contingency tables?
3. When would you use a Mann-Whitney *U* test or a Wilcoxon signed ranks test?

SPECIAL TOPICS IN INFERENTIAL STATISTICS

The application of the appropriate inferential statistic may appear simple and straightforward. However, several factors must be considered, beyond whether to apply a parametric or nonparametric statistic, when using any inferential statistic. This section discusses some special topics to consider when deciding on a strategy to evaluate data statistically.

Power of a Statistical Test

Inferential statistics are designed to help you determine the validity of the null hypothesis. Consequently, you want your statistics to detect differences in your data that are inconsistent with the null hypothesis. The **power** of a statistical test is its ability to detect these differences. Put in statistical terms, *power* is a statistic's ability to correctly reject the null hypothesis (Gravetter & Wallnau, 2010). A powerful statistic is more likely to detect the effects of your independent variables when they are present.

The issue of the power of your statistical test is an important one. Rejection of the null hypothesis implies that your independent variable affected your dependent variable. Failure to reject the null hypothesis may lead you to abandon a potentially fruitful line of research. Consequently, you want to be reasonably sure that your failure to reject the null hypothesis is not caused by a lack of power in your statistical test.

The power of your statistical test is affected by your chosen alpha level, the size of your sample, whether you use a one-tailed or two-tailed test, and the size of the effect produced by your independent variable.

Alpha Level As you reduce your alpha level (e.g., from .05 to .01), you reduce the probability of making a Type I error. Adopting a more conservative alpha level makes it more difficult to reject the null hypothesis. Unfortunately, it also reduces power. Given a constant error variance, a larger difference between means is required to obtain statistical significance with a more conservative alpha level (e.g., .01 instead of .05).

Sample Size The power of your statistical test increases with the size of your sample because larger samples provide more stable estimates of population parameters. In particular, the standard errors of the means from your treatments will be lower, so the likely positions of the population means fall within narrower bounds. Consequently, it is easier to detect small differences in population means and thus to reject the null hypothesis when it is false.

One-Tailed Versus Two-Tailed Tests A two-tailed test is less powerful than a one-tailed test. This can be easily demonstrated by looking at the critical values of t found in Table 2 in the Appendix. At 20 degrees of freedom, the critical value at $\alpha = .05$ for a one-tailed test is 1.73. For a two-tailed test, the critical value is 2.09. It is thus easier to reject the null hypothesis with the one-tailed test than with the two-tailed test.

Effect Size The degree to which the manipulation of your independent variable changes the value of the dependent variable is termed the **effect size**. To facilitate comparison across variables and experiments, effect size is usually reported as a proportion of the variation in scores *within* the treatments under comparison; for example, the effect size for the difference between two treatment means might be reported as $(M_2 - M_1)/s$, where s is the pooled sample standard deviation (Cohen, 1988). Measured in this way, effect size estimates the amount of overlap between the two population distributions from which the samples were drawn. Large effect sizes indicate relatively little overlap: The mean of Population 2 lies far into one tail of the distribution of Population 1, so a real difference in population means is likely to be detected in the inferential test (good power). Small effect sizes indicate great overlap in the population distributions and thus, everything else being equal, relatively little power. However, because inferential tests rely on the sampling distribution of the test statistic rather than the population distributions, you may be able to improve power in such cases by, for example, increasing the sample size.

In the past, effect sizes were reported rarely, but a growing recognition of their importance in the interpretation of data has led to a dramatic change in publication practices. In fact, according to the *Publication Manual of the American Psychological Association* (6th ed.), "For the reader to appreciate the magnitude or importance of a study's findings, it is almost always necessary to include some measure of effect size in the Results section" (APA, 2010, p. 34).

Determining Power Because the business of inferential statistics is to allow you to decide whether or not to reject the null hypothesis, the issue of power is important. You want to be reasonably sure that your decision is correct. Failure to achieve statistical significance in your experiment (thus not rejecting the null hypothesis) can be caused by many factors. Your independent variable actually may have no effect, or your experiment may have been carried out so poorly that the effect was buried in error variance. Or maybe your statistic simply was not powerful enough to detect the difference, or you did not use enough subjects.

Although alpha (the probability of rejecting the null hypothesis when it is true) can be set directly, it is not so easy to determine what the power of your analysis will be. However, you can work backward from a desired amount of power to estimate the sample sizes required for a study. To calculate these estimates, you must be willing to state the amount of power required, the magnitude of the difference that you expect to find in your experiment, and the expected error variance.

The expected difference between means and the expected error variance can be estimated from pilot research, from theory, or from previous research in your area. For example, if previous research has found a small effect of your independent variable (e.g., 2 points), you can use this as an estimate of the size of your effect.

Unfortunately, the proper amount of power is not easy to establish. There is no agreed-on acceptable or desirable level of power (Keppel, 1982). If you are willing and able to specify the values mentioned, however, you can estimate the size of the sample needed to detect differences of a given magnitude in your research. (See Gravetter & Wallnau, 2010, or Keppel, 1982, for a discussion on how to estimate the required sample size.)

Too much power can be as bad as too little. If you ran enough subjects, you could conceivably find statistical significance in even the most minute and trivial of differences. Similarly, when you use a correlation, you can achieve statistical significance even with small correlations if you include enough subjects. Consequently, your sample should be large enough to be sensitive to differences between treatments but not so large as to produce significant but trivial results.

The possibility of your results being statistically significant and yet trivial may seem strange to you. If so, the next section may clarify this concept.

Statistical Versus Practical Significance

To say that results are significant (statistically speaking) merely indicates that the observed differences between sample means are probably reliable, not the result of chance. Confusion arises when you give the word *significant* its more common meaning. Something "significant" in this more common sense is important or worthy of note.

The fact that the treatment means of your experiment differ significantly may or may not be important. If the difference is predicted by a particular theory and not by others, then the finding may be important because it supports that theory over the others. The finding also may be important if it shows that one variable strongly affects another. Such findings may have practical implications by demonstrating, for example, the superiority of a new therapeutic technique. In such cases, a statistically significant (i.e., reliable) finding also may have *practical significance*.

Advertisers sometimes purposely blur the distinction between statistical and practical significance. A number of years ago, Bayer aspirin announced the results of a "hospital study on pain other than headache." Evidently, groups of hospital patients were treated with Bayer aspirin and with several other brands. The advertisement glossed over the details of the study, but apparently the patients were asked to rate the severity of their pain at some point after taking Bayer or Brand X (the identities of both brands were probably concealed). According to the ad, "the results were significant—Bayer was better." However, the ad did not say in what way the results were significant. Evidently, the results were statistically significant and thus unlikely given that only chance was operating to produce the difference. Without any information about the pain ratings, however, you do not know if this finding has any practical significance. It may be that the Bayer and Brand X group ratings differed by less than 1 point on a 10-point scale. Although this average difference may have been reliable, it also may be the case that no individual could tell the difference between two pains so close together on the scale. In that case, the statistically significant difference would have no practical significance and would provide no reason for choosing Bayer over other brands of aspirin.

The Meaning of the Level of Significance

In the behavioral sciences, an alpha level of .05 (or 1 chance in 20) is usually considered the maximum acceptable rate for Type I errors. This level provides reasonable protection against Type I errors while also maintaining a reasonable level of power for most analyses. Of course, if you want to guard more strongly against Type I errors, you can adopt a more stringent alpha level, such as the .01 level (1 chance in 100).

Whatever alpha level that you determine is reasonable for your purposes, remember that this number does nothing more than provide a criterion for deciding whether the differences you have obtained are reliable. A difference is either reliable or it is not. If your results are significant at the .0001 level, they are not any more reliable than if they were significant at the .05 level. It does not mean your results are "more significant" or "more reliable" than significant results obtained at the .05 level. If the results are statistically significant at your chosen alpha level, it simply means you are willing to believe that the differences are real. However, lower alpha levels (moving from .05 to .01) allow you greater confidence in your decision about your results.

The importance of Type I errors may vary depending on the type of research and the purposes to which the information may be put. For example, applied research may be better evaluated at a less conservative alpha level (for example, $p < .10$). If you were testing the effectiveness of a new form of judicial instruction on the reduction of bias against Black defendants, a Type II error might be more serious than a Type I error. If you retain the null hypothesis when it is false, more Black defendants may be convicted as a result.

Ultimately, it is up to you to decide on an appropriate balance between Type I and Type II errors. Unfortunately, most journals will not publish a finding unless it is significant at least at the $p < .05$ level. Chapter 3 examined this issue in the discussion of publication practices.

Data Transformations

Sometimes you may find it necessary to transform your data with an appropriate **data transformation**. Transforming data means converting your original data to a new scale. For example, a simple transformation can be accomplished by adding or subtracting a constant to or from your data. You might do this if the original numbers are very large. When you compute some statistics, large numbers make the computations difficult. Subtracting a constant from each score can make the numbers manageable without affecting the relationships within the data. Conversely, adding a constant to each score might remove negative numbers.

When you add or subtract a constant, the shape of the original frequency distribution does not change. The mean of the distribution changes, but its standard deviation does not. When you multiply or divide by a constant, both the mean and standard deviation change. Such transformations, called *linear transformations*, simply change the magnitude of the numbers representing your data, but they do not change the scale of measurement.

Certain statistics can be used only if your data meet certain assumptions. If your data do not meet these assumptions, you could choose a different statistic. Unfortunately, this is not always desirable or possible. A nonparametric statistic that can be substituted for a parametric statistic may not exist for your particular situation. Another solution is to consider using a data transformation that will tend to correct the problem (e.g., by changing a skewed distribution of scores into a normal one or by removing heterogeneity of variance). Different problems with the data require different transformations to correct them. Table 14-7 lists some of the more popular data transformations and the conditions under which each might be used.

Data transformations to make data conform to the assumptions of a statistic are being used less and less frequently (Keppel, 1973). ANOVA, perhaps the most commonly used inferential statistic, appears to be very robust against even moderately

TABLE 14-7 **Data Transformations and Uses**

TRANSFORMATION	FORMULA	USE
Square root	$X' = \sqrt{X}$ *or* $X' = \sqrt{X + 1}$[a]	When cell means and variances are related, this transformation makes variances more homogeneous; also, if data show a moderate positive skew
Arcsin	$X' = 2 \arcsin \sqrt{X}$ *or* $X' = 2 \arcsin \sqrt{X \pm (1/2n)}$[b]	When basic observations are proportions and have a binomial distribution
Log	$X' = \log X$ *or* $X' = \log (X + 1)$[c]	Normalizes data with severe positive skew

[a] Formula used if basic observations are frequencies or if values of X are small.

[b] Formula used if values of X are close to 0 or 1.

[c] Formula used if value of X is equal to or near 0.

SOURCE: Information summarized from Tabachnick and Fidell, 2006, and Winer, 1971.

serious violations of its assumptions underlying the test. For example, Winer (1971) has demonstrated that even if the within-cell variances vary by a 3:1 ratio, the *F* test is not seriously biased. Transformations of the data may not be necessary in these cases. Also, when you transform your data, your conclusions must be based on the transformed scale and not the original. In most cases, this is not a problem. However, Keppel (1973) provides an example in which a square root transformation changed significantly the relationship between two means. Prior to transformation, the mean for Group 1 was lower than the mean for Group 2. The opposite was true after transformation.

Use data transformations only when absolutely necessary because they can be tricky. Sometimes transformations of data correct one aspect of the data (such as restoring normality) but induce new violations of assumptions (such as heterogeneity of variance). If you must use a data transformation, before going forward with your analysis, check to be sure that the transformation had the intended effect.

Alternatives to Inferential Statistics

Inferential statistics are tools to help you make a decision about the null hypothesis. Essentially, inferential statistics provide you with a way to test the reliability of a single experiment. When you reject the null hypothesis at $p < .05$, it means that a chance difference as large as or larger than the one obtained would occur only once (on the average) in 20 replications of the experiment.

Because such chance differences are relatively rare, you conclude that the difference you obtained was probably not due to chance but rather to the effect of the independent variable. If, in fact, the independent variable was the cause of the observed differences, then you would expect to obtain similar results on replication of the experiment. In other words, you would expect your findings to be reliable.

Inferential statistics cannot always be applied to assess the reliability of your results. You may have too few subjects (such as in single-subject or small-*n* research designs). Or you may have data that badly violate the assumptions of parametric tests with no appropriate nonparametric statistic to use instead. In these cases, you may test the reliability of your data by replication.

Replication means that you repeat your experiment. If your data are reliable, you should find a highly similar pattern of results after each replication. Replication does not mean that you have to conduct exactly the same experiment each time. Often a subsequent experiment in a series will include conditions that replicate those of the original experiment. The subsequent experiment may include conditions designed to test the effects of changing some parameters within the original context. The new experiment will provide a check on the original findings while providing new information.

Keep in mind that replication is not limited to small-*n* designs or situations in which violations of assumptions occur. You can include an element of replication in just about any study. Moreover, you need not limit yourself to replicating your own findings. If previous research shows a certain effect, you may wish to replicate that finding in your own research before extending your observations to new situations. Indeed, such replications are the heart of the scientific method. When successful, they demonstrate the reliability of findings both within the original context and across experimenters, subjects, and laboratories. When unsuccessful, they point to potentially important variables that may limit the generality of findings to particular situations and parameters. Either result can be important for the advancement of scientific knowledge.

Inferential statistics were developed to assess the reliability of findings within the confines of a single set of observations. By providing an index of probable reliability, they reduce the need for direct replication and thus save time and money by reducing the requirement for subjects. Nevertheless, they should not be viewed as a substitute for replication. Probably no finding in psychology has been accepted on the basis of a single experiment that was statistically significant at some alpha level. The value of inferential statistics is found not so much in the elimination of replication as in their warning that the effects apparent in research data may result from nothing more than random factors. Human beings are extremely good at recognizing patterns even when such patterns are simply the result of "noise." Inferential statistics can control the human tendency to interpret every apparent trend or difference in the data as if it were meaningful.

Inferential statistics sometimes may lack sufficient power and therefore may fail to detect effects that are clearly shown by replication. A case in point is provided by a series of experiments conducted by one of the authors of this text (Abbott) to test the effect of predictable versus unpredictable shock schedules on pain sensitivity. In each experiment, three groups of eight rats were exposed to a schedule of predictable shock, unpredictable shock, or no shock. The subjects were then tested in the same apparatus for pain sensitivity by means of the "tail-flick" test. In the tail-flick test, a hot beam of light was focused on the rat's tail. The length of time elapsing until the rat flicked its tail out of the beam (a protective reflex) indicated the degree of pain sensitivity.

The results of the first experiment indicated that the two groups exposed to shock were less sensitive to the heat than the group exposed to no shock, replicating a well-established finding. In addition, the group exposed to unpredictable shock seemed less sensitive than the group exposed to predictable shock. However, this effect was not statistically significant ($p < .05$).

Parameters of the experiment were twice altered in ways that were expected to increase the size of the predictability effect (if it existed), and the experiment was replicated. However, each replication produced virtually the identical result. On each occasion, the unpredictable shock group demonstrated less sensitivity to pain than the predictable shock group, and each time this difference was not statistically significant.

The problem could be dealt with by taking measures to increase the power of the statistical test (such as increasing sample sizes or going to a matched groups design). However, to do so would appear to be a waste of resources. In this case, the reliability of the finding was already established through replication even though the statistical analysis itself indicated that the results were probably not reliable.

Inferential statistics are simply a guide to decision making and are not the goal of the research project. As such, you should not design and conduct your research in a particular way simply because a particular inferential statistic is available to analyze such a design. Much like designing your experiment before developing hypotheses, choosing a statistical test before designing a study can place unwanted restrictions on your research. For example, you may not be able to manipulate your independent variable the way you would like and may miss some important relationships. Instead, design your study to answer your research questions in the clearest way possible and then select the method of analysis (whether inferential statistic or replication) that works best for that design.

QUESTIONS TO PONDER

1. What is an effect size, and why is it important to include some measure of effect size along with the results of your statistical test?
2. What is meant by the power of a statistical test, and what factors can affect it?
3. Does a statistically significant finding always have practical significance? Why or why not?
4. When are data transformations used, and what should you consider when using one?
5. What are the alternatives to inferential statistics for evaluating the reliability of data?

SUMMARY

This chapter has reviewed some of the basics of inferential statistics. Inferential statistics go beyond simple description of results. They allow you to determine whether the differences observed in your sample are reliable. Inferential statistics allow you to make a decision about the viability of the null hypothesis (which states that there is no difference among treatments) while controlling the probability of rejecting the null hypothesis when it is in fact true (Type I error). The two types of inferential statistics are parametric and nonparametric. Parametric tests (such as the *t* test and ANOVA) make assumptions about the populations underlying your samples. For example, these tests assume that the sampling distribution of means is normal and that there is homogeneity of within-cell variances. Parametric statistics are designed for use when your data are scaled on at least an interval scale. If your data seriously violate the assumptions of a parametric test or your data are scaled on a nominal or ordinal scale, a nonparametric statistic can be used (such as chi-square or the Mann–Whitney *U* test). These tests are usually easier to compute than parametric tests. However, they are less powerful and more limited in application. Nonparametric statistics may not be available for higher-order factorial designs.

Statistical significance indicates that the difference between your means was unlikely if only chance were at work. It suggests that your independent variable had an effect. Two factors contribute to a statistically significant effect: the size of the difference between means and the variability among the scores. You can have a large difference between means, but if the variability is high, you may not find statistical significance. Conversely, you may have a very small difference and find a significant effect if the variability is low. Measures of effect size can help to assess how strong your treatment differences are relative to the within-treatment variability of your scores. Most journals now insist that some measure of effect size be included along with the *p*-values from your statistical analysis.

Consider the power of your statistical test when evaluating your results. If you do not find statistical significance, perhaps no differences exist. Or it could mean that your test was not sensitive enough to pick up small differences that do exist. Sample size is an important contributor to power. Generally, the larger the sample, the more powerful the statistic. This is because larger samples are more representative of the underlying populations than are small samples. Use a sample that is large enough to be sensitive to differences but not so large as to be oversensitive. There are methods for determining optimal sample sizes for a given level of power. However, you must be willing and able to specify an expected magnitude of the treatment effect, an estimate of error variance, and the desired power. The first two can be estimated from pilot data or previous research. Unfortunately, there is no agreed-on acceptable level of power.

An alpha level of .05 is the largest generally acceptable level for Type I errors. This value has been chosen because it represents a reasonable compromise between Type I and Type II errors. In some cases (such as in applied research), the .05 level may be too conservative. However, journals probably will not publish results that fail to reach the conventional level of significance.

Data transformations are available for those situations in which your data are in some way abnormal. You may transform data if the numbers are large and unmanageable or if your data do not meet the assumptions of a statistical test. The transformation of data to meet assumptions of a test, however, is being done less frequently because inferential statistics tend to be robust against the effects of even moderately severe violations of assumptions. Transformations should be used sparingly because they change the nature of the variables of your study.

Inferential statistics are not the only means available for assessing the reliability of your findings. Where samples are necessarily small or an appropriate inferential statistic is not available for your research design, don't forget that you can still assess reliability by actual replication of your study.

KEY TERMS

inferential statistics
standard error of the mean
degrees of freedom (*df*)
Type I error
Type II error
alpha level (α)
critical region
t test
t test for independent samples
t test for correlated samples
z test for the difference between two proportions
analysis of variance (ANOVA)
F ratio
p value
planned comparisons
unplanned comparisons
per-comparison error
familywise error
analysis of covariance (ANCOVA)
chi-square (χ^2)
Mann–Whitney *U* test
Wilcoxon signed ranks test
power
effect size
data transformation

15 CHAPTER

CHAPTER OUTLINE

Correlational and Experimental Multivariate Designs
- Correlational Multivariate Design
- Experimental Multivariate Design
- Causal Inference

Assumptions and Requirements of Multivariate Statistics
- Linearity
- Outliers
- Normality and Homoscedasticity
- Multicollinearity
- Error of Measurement
- Sample Size

Correlational Multivariate Statistical Tests
- Factor Analysis
- Partial and Part Correlations
- Multiple Regression
- Discriminant Analysis
- Canonical Correlation

Experimental Multivariate Statistical Tests
- Multivariate Analysis of Variance
- Multiway Frequency Analysis

Multivariate Statistical Techniques and Causal Modeling
- Path Analysis
- Structural Equation Modeling

Multivariate Analysis: A Cautionary Note

Summary

Key Terms

Using Multivariate Design and Analysis

During discussions of experimental and nonexperimental design, previous chapters assumed that only one dependent variable was included in a design or that multiple dependent variables were treated separately in any statistical tests. This approach to analysis is called a **univariate strategy**. Although many research questions can be addressed with a univariate strategy, others are best addressed by considering dependent variables together in a single analysis. When you include two or more dependent measures in a single analysis, you are using a **multivariate strategy**.

This chapter introduces the major multivariate analysis techniques. Keep in mind that providing an in-depth introduction to these techniques in the confines of one chapter is impossible. Such a task is better suited to an entire book. Also, the complex and laborious calculations needed to compute multivariate statistics are better left to computers. Consequently, this chapter does not discuss the mathematics behind these statistical tests except for those cases in which some mathematical analysis is required to understand the issues. Instead, this chapter focuses on practical issues: applications of the various statistics, the assumptions that must be met, and interpretation of results. If you want to use any of the statistics discussed in this chapter, read *Using Multivariate Statistics* (Tabachnick & Fidell, 2001, 2006) or one of the many monographs published by Sage Publications (such as Asher, 1976, or Levine, 1977).

CORRELATIONAL AND EXPERIMENTAL MULTIVARIATE DESIGNS

A **multivariate design** is a research design in which you include multiple dependent or multiple predictor and/or criterion variables. Analysis of data from such designs requires special statistical procedures.

Multivariate design and analysis apply to both experimental and correlational research studies. The following sections describe some of the available multivariate statistical tests.

Correlational Multivariate Design

If you include multiple measures in a correlational study, you could calculate separate bivariate correlations (e.g., Pearson correlations) for all possible pairs of those measures. Or you could use a *multivariate correlational analysis* in which you include all your measures in a single analysis.

Several multivariate analyses are designed to assess complex correlational relationships among multiple dependent variables. For example, the goal of **multiple regression** is to explain the variation in one variable (the dependent or *criterion variable*) based on variation in a set of others (the *predictor variables*). What constitutes a "predictor variable" and a "criterion variable" is not related to anything inherent in the variable itself. Rather, you decide which variables to use as predictors based on your research question. Relevant previous research, theory, or practical experience should guide your decision about which variables should be measured and what role each variable should play in your analysis.

Two other multivariate techniques used to evaluate relationships in a correlational study are discriminant analysis and canonical correlation. **Discriminant analysis** is a variation of multiple regression in which your criterion variable is measured nominally (e.g., yes/no). **Canonical correlation** allows you to evaluate the relationship between two *sets* of variables, one of which may be identified as a predictor variable set and the other as the criterion variable set.

In some research situations (e.g., questionnaire and test construction), you may reduce a large set of variables to smaller sets that consist of variables relating to one another. Factor analysis is used for this purpose. In **factor analysis**, several dependent variables are analyzed to find out if any of them share common underlying dimensions called *factors*. You examine the dependent variables that make up the factors to identify the dimension that those factors represent.

Advantages of the Correlational Multivariate Strategy Single multivariate analysis has two major advantages over multiple bivariate analyses. First, if you conduct a large number of independent bivariate correlation analyses, you increase your risk of finding relationships that occur merely by chance. Multivariate statistics allow you to look at complex relationships while controlling such statistical errors. Second, independent univariate analyses allow you to evaluate relationships only between *pairs* of variables. You may discover that Variable *X* correlates highly with Variable *Y*. However, what you do not know is whether this high correlation will persist when you consider a third variable, *Z*. It may be that *X* is correlated with *Y* only because *Z* is highly correlated with *X*. Multivariate statistics provide the information needed to evaluate the importance of a predictor variable for explaining variability in the criterion variable, given the effects of other predictor variables.

Experimental Multivariate Design

The logic of univariate experimental design applies to multivariate design. That is, you manipulate one or more independent variable(s) and look for changes in the values of your dependent variables. The major difference between a univariate experimental strategy and a multivariate experimental strategy is how dependent variables are handled. When you use a univariate strategy, multiple dependent measures are analyzed separately with multiple statistical tests. In contrast, when you use a multivariate strategy, multiple dependent variables are combined statistically (based on the correlations among them) and analyzed with a single statistical test.

Implied in your choice of a multivariate design over a univariate design is that your dependent measures are correlated. Typically, you include multiple dependent measures because you have some reason to believe that those measures are important to the phenomenon under study and that those measures relate in some way to one another. Multivariate statistical techniques take into account the correlations among your dependent measures and, in most cases, use them to your advantage.

Multivariate Statistical Tests for Experimental Designs The two multivariate statistics most widely used to analyze multiple dependent variables in an experimental design are **multivariate analysis of variance (MANOVA)** and *multivariate analysis of covariance (MANCOVA)*. Another multivariate statistic that is commonly used when your dependent variable is categorical (e.g., guilty or not guilty) is *multiway frequency analysis*. As with univariate statistics, these tests help you evaluate the reliability of the relationship between your independent variable (or variables) and your dependent variables.

Advantages of the Experimental Multivariate Strategy A multivariate experimental strategy has several advantages over a univariate strategy. First, collecting several dependent measures and treating them as a correlated set may reveal relationships that might be missed if a traditional univariate approach were taken. Because multivariate statistical tests consider the correlations among dependent variables, they tend to be more powerful than separate univariate tests of those same dependent variables. Second, because all your dependent variables are handled in a single analysis, complex relationships among variables can be studied with less chance of making a Type I error than when using multiple univariate tests (Bray & Maxwell, 1982).

A third advantage of the multivariate strategy is realized when you have used a within-subjects design. A fairly restrictive set of assumptions underlies the univariate within-subjects ANOVA that are often difficult to satisfy. Using MANOVA allows you to analyze your data with less concern over these restrictive assumptions.

Causal Inference

Multivariate techniques allow you to draw some *tentative* causal inferences from your correlational data. At the least, when properly applied, they allow you to have greater confidence in possible causal connections among variables. However, remember that you are still using correlational data, so any causal inferences you draw must be discussed with caution.

Path analysis and *structural equation modeling* are two techniques that allow you to explore causal models relating your variables. Path analysis, which applies multiple regression analysis to the investigation of possible causal relationships among variables, begins with a theory or model specifying a causal chain of events involving several variables and a behavior. For example, a model may suggest that a consumer's buying behavior will not occur until the consumer first finds out about a product, then generates positive ideas about the product, and finally forms an intention to buy the product. You could obtain measures on the degree to which the consumer was familiar with the product, had positive ideas about it, and intended to buy it. You could then enter the measures into a series of multiple regression analyses. Based on the results, you could test the validity of your theory or model and begin to form some tentative conclusions about possible causal relationships among your variables. Structural equation modeling is related to path analysis and allows you to more completely explore potential causal models linking your variables.

ASSUMPTIONS AND REQUIREMENTS OF MULTIVARIATE STATISTICS

Before using multivariate statistics, you must check to see that your data meet the assumptions and requirements underlying the statistic to be used. These assumptions include linearity, normality, and homoscedasticity. In addition, you must evaluate your data for the presence of outliers and measurement error and for sufficient sample size.

Linearity

An assumption underlying bivariate correlational statistics is that the relationship between continuously measured variables is linear (see Chapter 13 for a more complete discussion). Violation of this assumption leads to an underestimation of the degree of relationship between variables. Multivariate statistics, which are all based on correlations (even MANOVA), also assume that the relationships among continuously measured variables are linear. You check for linearity by visually inspecting scatter plots of pairs of variables. If your data are linear, then all the points should follow a straight line. Nonlinear data, in contrast, will show a horseshoe-shaped function (Tabachnick & Fidell, 2001).

Whereas mild deviations from linearity probably will not lead to a serious underestimation of a relationship by multivariate statistics, moderate to serious deviations may. If your data are nonlinear, you may be able to correct the problem by transforming your data. You may have to transform both offending variables in order to restore linearity. After any transformation, you should again inspect the scatter plots to see if the transformation had its intended effect.

Outliers

Bivariate correlational statistics work by fitting the best straight line to the data. This *regression line* minimizes the distance of the data points from the line according to some statistical criterion (usually a least-squares criterion). If your data set has extreme scores, or *outliers*, how the regression line fits the data may not represent the trend shown by the majority of scores. Outliers change the slope of the regression line calculated from your data. They also affect both the magnitude and the sign of the calculated correlation. (See Chapter 13 for a more complete discussion.)

Identifying Outliers Two types of outliers that must be considered in multivariate statistics are univariate outliers and multivariate outliers. A *univariate outlier* is a deviant score on one measure from a given source (e.g., from a single subject), whereas a *multivariate outlier* is a deviant score on a combination of variables from a single source.

Univariate outliers may be detected by converting raw scores to z scores. If the z score is very deviant (such as 3), then that raw score is considered to be a univariate outlier, especially with a large sample (Tabachnick & Fidell, 2001). Another way to look for outliers is to evaluate the amount of skewness in your data. If your data are skewed, then outliers probably exist. However, note that measures of skewness (see Chapter 13) detect skewness if outliers exist only in one tail of your distribution.

Detecting multivariate outliers is more difficult than detecting univariate outliers. Multivariate outliers can be detected either statistically or graphically (Tabachnick & Fidell, 2001). Using the statistical method, you obtain a statistic called the *Mahalanobis Distance* from a statistical program such as SPSS. The Mahalanobis Distance represents the distance between a particular case and the centroid (the point created by the means of all the variables in the analysis) of remaining cases. The statistical program also calculates a discriminant analysis that tells you which case separates from the other cases in the analysis (Tabachnick & Fidell, 2001). A multivariate outlier with an unusual combination of scores will be weighted heavily in the discriminant function equation, yielding a significant Mahalanobis Distance of the outlier from other cases. There are other statistical techniques that you can use to detect multivariate outliers (see, e.g., Tabachnick & Fidell, 2001, pp. 68–70). You also can detect multivariate outliers by inspecting plots of residuals provided by multiple regression programs. Any point on the plot that is distant from other points is a multivariate outlier. Screening for multivariate outliers should be done again after the data are cleaned up because some outliers may “hide” behind others (Tabachnick & Fidell, 2001).

Dealing With Outliers You can use several strategies to deal with outliers if you discover them in your data. To normalize the distribution, Tabachnick and Fidell (2001) suggest using one of several data transformations on the offending variable. Data with a moderate positive skew should be transformed with a square root transformation. You should use a logarithmic transformation if your data have a more serious positive skew. Again, transformations such as these reduce the impact of outliers if they are found only in one tail of your distribution. If outliers exist in both tails, transformations may not help (Tabachnick & Fidell, 2001).

If your data are negatively skewed, you use a *reflecting strategy*. The first step in reflecting is to transform your data so that they are positively skewed. You accomplish this by subtracting each score from the highest score in the distribution and adding 1. The resulting positively skewed data are then transformed with either a square root or log transformation, depending on the degree of skewness.

Another way to deal with outliers is to delete from the analysis either all data from the subject with the outlying scores or the entire variable. The disadvantage to this procedure is that you lose data. If you start with a relatively small sample, the loss of data may preclude using multivariate statistics.

Finally, you should check for transcription and other data-entry errors. Sometimes outliers are caused by entering the wrong numbers or by telling the computer to look for data in the wrong positions. Any erroneous data should be corrected.

Of all the requirements of multivariate statistics, detecting outliers is probably the most important. The presence of just a single multivariate outlier can change the results of your analysis and affect your conclusions. Consequently, you should check and correct for both univariate and multivariate outliers.

Normality and Homoscedasticity

As is the case with bivariate statistics, multivariate statistics assume that the population distribution underlying your sample distribution is normal. This is the assumption of *normality*. Transform skewed data with one of the indicated transformations, in order to normalize the distributions, before using any multivariate statistic.

Homoscedasticity is related to normality and is the assumption that "the variability in scores on one variable is roughly the same at all values of the other variable" (Tabachnick & Fidell, 2001, p. 79). Figure 15-1 shows two scatter plots of two hypothetical variables. Panel (a) shows the pattern of data indicating homoscedasticity. Notice that the shape of the scatter plot created by the data points is elliptical. If both variables are normally distributed, homoscedasticity results.

Contrast the scatter plot shown in panel (b) of Figure 15-1 with the one shown in panel (a). Notice how the shape of the scatter plot has changed from elliptical to conical. The conical pattern of data points indicates that *heteroscedasticity* is present. Heteroscedasticity usually occurs because the distribution of one or more variable(s) included in the analysis is skewed. To eliminate heteroscedasticity, apply one of the data transformations previously discussed.

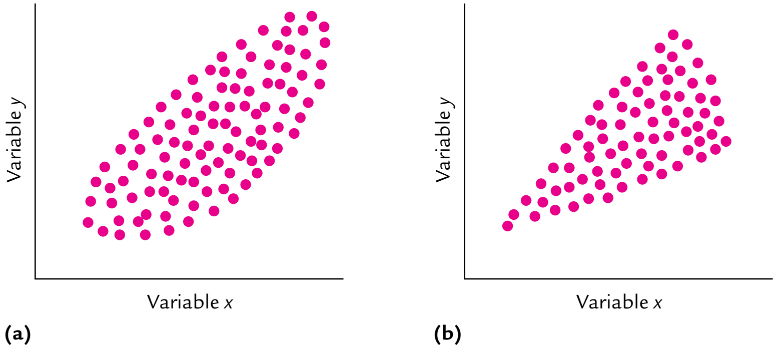


FIGURE 15-1 Homoscedasticity and heteroscedasticity: (a) homoscedasticity between two variables and (b) heteroscedasticity between two variables.

Multicollinearity

Multicollinearity results when variables in your analysis are highly correlated (Tabachnick & Fidell, 2001). The impact of multicollinearity is complex and beyond the scope of this chapter. If two variables are highly correlated, one of them should be eliminated from the analysis. The high correlation means the two variables are measuring essentially the same thing, so little is lost by eliminating one of them.

Error of Measurement

The heart of the research process is identifying important variables to study, measuring those variables, establishing relationships among variables, and drawing conclusions about behavior based on those relationships. Drawing valid conclusions about the behavior under study requires that your variables be accurately measured. Inaccurate measurement may lead to an inordinate number of Type II errors. For example, you may conclude that a theoretical model is invalid when the reason for rejecting the model was in fact an inaccurate measurement of variables.

In a perfectly ordered world with perfect measuring devices, you could obtain the *true value* of your dependent variable. Unfortunately, we do not live in a perfectly ordered world, nor do we have perfect measuring devices. Consequently, the best you can do is to *estimate* the true value of a variable by obtaining an *observed value*. The difference between the true value of a variable and your observed value is the *error of measurement* (or simply, *measurement error*). Figure 15-2 shows the relationship between a variable's true value, observed value, and measurement error. Notice that the observed value is a function of both the true value of the variable and the measurement error (Asher, 1976).

Error of measurement is a problem for both multivariate and univariate research. It is particularly troublesome when you adopt a multivariate strategy because it leads to an underestimation of the correlations among variables that are used to compute the various multivariate statistics (Asher, 1976; Hunter, 1987). This leads to Type II errors.

Measurement error can arise from many sources, including incomplete, inaccurate, or biased sources of information. For example, if you were interested in studying the relationship between three predictor variables (gender, socioeconomic status, and education level) and crime rate, you need a good source for each of these four variables. Consider the crime rate. You could obtain records of crimes reported to

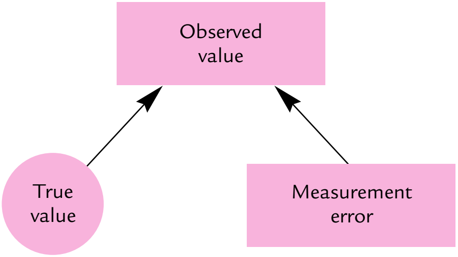


FIGURE 15-2 The observed value of any variable (top) is a function of the true value of that variable and the measurement error (bottom).

the police, but this source may not be complete because many crimes go unreported. The best way to avoid this source of measurement error is to use multiple sources of information.

Another source of measurement error is inaccurate or invalid measurement devices. Defects in mechanical recording devices, poorly designed rating scales, and the like all contribute to measurement error. To avoid this source of error, be sure that your equipment is in working order and that you have adequately pretested your measures.

Sample Size

Fairly large sample sizes are needed for multivariate analyses. The large sample size is necessary because the correlations used to calculate these statistics are not very stable when based on small samples. A multivariate analysis that uses a small sample may result in an unacceptable Type II error rate. This occurs because unstable correlations tend to provide less reliable estimates of the degree of relationship among your variables.

Tabachnick and Fidell (2001, p. 117) offer the following formula for computing the sample size required for a multiple regression analysis:

$$N \geq 50 + 8m$$

where m = the number of predictor variables. So, if you have five predictor variables, you would need a minimum of 90 participants in your sample. Larger samples may be needed if your data are skewed, there is substantial measurement error, or you anticipate weak relationships among variables (Tabachnick & Fidell, 2001). Tabachnick and Fidell also caution that you can have *too large* a sample. With overly large samples, very weak relationships that may have neither theoretical nor practical value can achieve statistical significance.

To summarize, several factors should be considered before using multivariate statistics. Make sure that your data meet the assumptions of the test you are going to use (i.e., normality, linearity, and homoscedasticity), that you have removed any outliers or minimized their effects through transformation, that you have considered error of measurement, and that you have gathered a sufficiently large sample. If you violate the assumptions of the test or fail to take into account the other important factors, the results that you obtain may not be valid.

QUESTIONS TO PONDER

1. What statistics are used to evaluate correlational and experimental multivariate relationships?
2. What are the key assumptions and requirements of multivariate statistics?
3. How do various violations of the assumptions underlying multivariate statistics affect your data analysis?

CORRELATIONAL MULTIVARIATE STATISTICAL TESTS

Now that you are familiar with the general logic behind multivariate statistics and understand the assumptions and requirements of these tests, we can explore some of the more popular multivariate data analysis techniques. This discussion begins with an examination of factor analysis, then examines the techniques used to analyze multivariate data from experimental designs (e.g., MANOVA and multiway frequency analysis), and, finally, examines techniques for causal modeling.

Factor Analysis

Imagine that you are interested in measuring the degree to which males conform to male social norms. Before you conduct your study, you need to find a way to define just what those norms are. While reviewing the literature, you discover that there are several male social norms that are relevant to male social behavior. You decide to design a questionnaire including 100 items to measure male social norms and administer it to a sample of male participants.

After all your participants have completed the questionnaire, you now face the task of determining the underlying nature of male social norms. One question that interests you is whether all the questions on your questionnaire measure a single dimension (such as aggressiveness) or several dimensions (such as aggressiveness, competitiveness, and dominance). Your search for the dimensions underlying male social norms lends itself perfectly to factor analysis.

Factor analysis operates by extracting as many significant factors from your data as possible, based on the bivariate correlations between your measures. A *factor* is a dimension that consists of any number of variables. In your study of male social norms, for example, you may find that your 100 questions actually measure three underlying dimensions (e.g., aggressiveness, competitiveness, and dominance). Factor analysis involves extracting one factor (such as aggressiveness) and then evaluating your data for the existence of additional factors.

The successive factors extracted in factor analysis are not of equal strength. Each successive factor accounts for less and less variance. Typically, the first two or three factors will be the strongest (i.e., account for the most variance). The strength of a factor is indicated by its *eigenvalue* (for a more complete discussion of eigenvalues, see Tatsuoka, 1971, or Tabachnick & Fidell, 2001). Factors with eigenvalues of less than 1.0 usually are not interpreted.

Factor Loadings To determine the dependent variables constituting a common factor, *factor loadings* are computed. Each factor loading is the correlation between a measure and the underlying factor. A positive factor loading means that a variable positively correlates with the underlying dimension extracted whereas a negative loading means that a negative correlation exists. By convention, loadings are interpreted only if they are equal to or exceed .30.

Rotation of Factors After you have obtained your factor loadings, you must interpret them. The factor loadings computed initially are often difficult to interpret because they are somewhat ambiguous. *Factor rotation* is used to make the factors distinct by maximizing high correlations and minimizing low correlations (Tabachnick & Fidell, 2001). Rotated factors will include more distinct clusters of factor loadings than unrotated factors and are thus easier to interpret. Two types of rotation are orthogonal rotation and oblique rotation.

In *orthogonal rotation*, the axes representing the factors remain perpendicular when rotated around fixed points representing your data. Orthogonal rotation assumes that your measures are uncorrelated and consequently that the factors extracted are uncorrelated. Generally, orthogonal rotation is preferred over oblique rotation because the results are easier to interpret. The most popular orthogonal rotation method is *varimax*, which maximizes the variance of loadings on each factor and simplifies factors (Tabachnick & Fidell, 2001).

In *oblique rotation*, the angle between the axes, as well as the orientation of the axes in space, may change. Oblique rotation assumes that your measures and factors are correlated. If you have good reason to believe that your measures are correlated, oblique rotation might be a better choice than orthogonal rotation.

Principal Components and Principal Factors Analysis Two types of factor analysis are *principal components analysis* and *principal factors analysis*. Panel (a) of Table 15-1 shows a standard three-variable correlation matrix. Remember, such correlations are used to calculate factor loadings. Panel (b) shows the same correlation matrix completed by filling in the correlations missing from the matrix in panel (a). Notice that the values on the diagonal of the matrix are all 1s. In principal components analysis, the diagonal of the completed correlation matrix is filled with 1s. In contrast, principal factors analysis completes the correlation matrix by entering *communalities*

TABLE 15-1 Two Correlation Matrices for Three Variables

(A) HYPOTHETICAL CORRELATION MATRIX

	Variable 1	*Variable 2*	*Variable 3*
Variable 1			
Variable 2	.71		
Variable 3	.61	.74	

(B) COMPLETED CORRELATION MATRIX

	Variable 1	*Variable 2*	*Variable 3*
Variable 1	1.00	.71	.61
Variable 2	.71	1.00	.74
Variable 3	.61	.74	1.00

along the diagonal. Essentially, communality is a measure of a variable's reliability and is fairly easy to obtain after factor analysis. In practice, however, you need these values before analysis. Various techniques have been proposed for estimating communalities (see Bennett & Bowers, 1976), none of which is much better than any other.

Your choice between principal components and principal factors analysis rests on the goals of the analysis. If your goal is to reduce a large number of variables down to a smaller set and to obtain an empirical summary of the data, then principal components analysis is most appropriate. If your research is driven by empirical or theoretical predictions, then principal factors analysis is best (Tabachnick & Fidell, 2001). In the absence of any clear information on which technique is best, you should probably use principal components in those situations in which you do not have any empirical or theoretical guidance on the values of the communalities.

Exploratory Versus Confirmatory Factor Analysis A distinction also is made between exploratory factor analysis and confirmatory factor analysis (Tabachnick & Fidell, 2001). *Exploratory factor analysis* is used when you have a large set of variables that you want to describe in simpler terms and you have no a priori ideas about which variables will cluster together. Exploratory factor analysis is often used in the early stages of research to identify the variables that cluster together. From such an analysis, research hypotheses can be generated and tested (Tabachnick & Fidell, 2001). *Confirmatory factor analysis* is used in later stages of research where you can specify how variables might relate given some underlying psychological process (Tabachnick & Fidell, 2001).

QUESTIONS TO PONDER

1. When is factor analysis used and what do factor loadings tell you?
2. Why are factors rotated in factor analysis?
3. What is the difference between principal components and principal factors analysis?
4. When do you use exploratory or confirmatory factor analysis?

Partial and Part Correlations

Sometimes two variables are both influenced by a third variable. If this third variable was not held constant when the data were collected, it can affect the apparent relationship between the two variables of interest. However, if you have recorded the values of the third variable along with the other two, you can statistically evaluate the impact of the third variable. Partial correlation and part correlation (also called the *semipartial correlation*) are two statistics that determine the correlation between two variables while statistically controlling for the effect of a third.

Partial Correlation **Partial correlation** allows you to examine the relationship between two variables with the effect of a third variable *removed* from both of these variables. For example, suppose you are interested in the factors relating to performance on the Scholastic Assessment Test (SAT). You obtain the SAT scores from

500 high school seniors, as well as their grade point averages (GPA). You also collect data on the parents' educational level (PE) in the belief that it may affect SAT scores. Specifically, you are interested in the relationship between GPA and SAT scores but are concerned that PE may confound the relationship between GPA and SAT. You want to look at the relationship between GPA and SAT with any effect of PE removed. This problem calls for a partial correlation.

Imagine that a causal relationship exists between PE and GPA. In that case, variations in PE would induce variations in GPA. Imagine that a causal relationship also exists between PE and SAT scores. In that case, variations in PE also would induce variations in SAT scores. If these were both direct relationships, then SAT scores and GPA would tend to rise and fall together as PE rose and fell. In other words, SAT scores and GPA would be positively correlated. This positive correlation would emerge even if there were no direct causal connection between SAT scores and GPA.

Would a correlation remain if you could somehow remove the common influence that PE has on SAT scores and GPA? This is what partial correlation attempts to determine, as shown in Figure 15-3. Panel (a) shows a scatter plot of the relationship between PE and GPA. The straight line through the points represents the trend relating changes in GPA to changes in PE. If you could remove this trend, the GPA scores would show less variability, and they would show no systematic change as PE changed.

In fact, you *can* remove this trend statistically by calculating a *residual score* for each data point. This residual score is the distance from the point to the line, measured (in this case) in terms of GPA. Panel (b) in Figure 15-3 plots the residual scores as a function of PE. Note that the trend relating PE to GPA is now flat. The changes in GPA induced by changes in PE have been statistically removed. The same process of statistically removing the effect of PE is also applied to the SAT scores. Panels (c) and (d) of Figure 15-3 show the relationship before and after this removal. Partial correlation is then determined by correlating the residual GPA scores, panel (b), with the residual SAT scores, panel (d). This is the correlation of GPA and SAT with the effect of PE removed from both.

Fortunately, there is an easier way to compute partial correlation than by graphing and subtracting. First, you find the simple correlations between your three variables. Then these correlation coefficients are entered into a special partial correlation formula.

Partial correlation is not limited to three-variable cases. You can examine the relationship between two variables with the effects of several others removed. You can learn more about these more complex partial correlations in Thorndike (1978).

Part Correlation

In some cases, you may want to examine the relationship between two variables when the influence of a third is removed from only *one* of these variables. **Part correlation** (also known as *semipartial correlation*) is used in this situation.

Conceptually, part correlation is similar to partial correlation. As with partial correlation, the relationship between one variable (such as SAT) and the variable to be removed (such as PE) is determined, and residual scores are calculated. This yields an SAT score for each participant with the effect of PE held constant (Thorndike, 1978). These residual scores are then correlated with the *raw* scores of the other variable (GPA) to yield the part correlation coefficient (Thorndike, 1978).

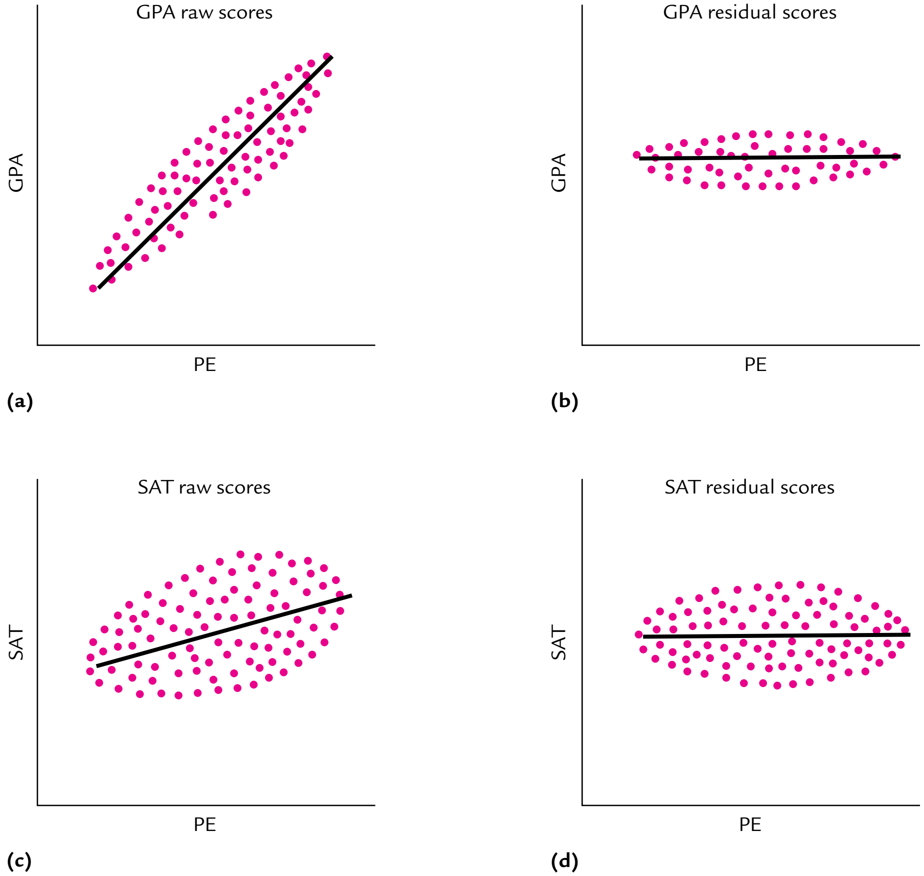


FIGURE 15-3 Four figures showing the logic behind partial correlation.

In practice, part correlation is computed by using a formula similar to that used for partial correlation.

QUESTIONS TO PONDER

1. What do partial and part correlations tell you?
2. When would you use partial correlation?
3. When would you use part correlation?

Multiple Regression

Assume that you are interested in studying the variables correlated with college students' attitudes toward seeking counseling for personal problems. You are interested in investigating those variables that relate to a student having either a positive or negative attitude toward seeking professional counseling. Multiple regression analysis is the best statistic to address such an issue. You will have a single measure of attitude toward counseling (the criterion or dependent variable) and several measures that might relate to that attitude (predictor variables).

The Multiple Regression Equation Chapter 13 discussed bivariate linear regression and provided the linear equation for that analysis. The logic developed for the bivariate case can be easily extended to the multivariate case. The linear equation for multiple regression is

$$\hat{Y} = b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + \text{constant}$$

where $\hat{Y}$ is the predicted criterion score; b_1, b_2, b_3, b_4, and b_5 are the regression weights associated with the predictors; X_1, X_2, X_3, X_4, and X_5 are the values of the predictors; and *constant* is the *y*-intercept.

Types of Regression Analysis The several types of regression analysis include simple, hierarchical, and stepwise analyses. The major difference between these types is how your predictor variables are entered into the regression equation, which may affect the regression solution.

In *simple regression analysis* (the type used in the example to follow), all predictor variables are entered together. Each predictor variable is assessed as if it had been entered after each of the other predictors had been entered (Tabachnick & Fidell, 2001). In *hierarchical regression*, you specify the order in which your variables are entered into the regression equation. You use hierarchical regression if you have a well-developed theory or model suggesting a certain causal order. In *stepwise regression*, the order in which variables are entered is based on a statistical decision, not on a theory.

When you enter variables into a stepwise regression analysis, the order in which predictors are entered is determined by the qualities of the sample data. The first variable entered is the one accounting for the most variance in the dependent measure. The next variable entered is the one that adds most to the ability of the regression equation to account for the variance in the dependent variable (i.e., increases *R*-square the most). Variables are entered one at a time until none of the remaining variables add significantly to *R*-square.

Your choice of regression strategies should be based on your research questions or underlying theory. If you have a theoretical model suggesting a particular order of entry, use hierarchical regression. In the absence of any well-specified theory, you should usually choose simple regression. Stepwise regression is used infrequently because it tends to capitalize on chance. Sampling and measurement error tend to make unstable correlations among variables in stepwise regression. Thus, the statistical decisions used to determine order of entry may vary considerably from sample to sample. The resulting regression equation may be unique to a particular sample.

Multiple R and R-Square **Multiple *R*** is the correlation between the predicted values of *Y* ($\hat{Y}$) and the observed values of *Y*. ***R*-square** is simply the square of multiple *R* and provides an index of the amount of variability in the dependent variable accounted for by the predictor variables (Roscoe, 1975). There is a problem with *R*-square. Because of sampling error, *R*-square tends to overestimate the variance accounted for, especially with small samples (Tabachnick & Fidell, 2001). *Adjusted R-square* compensates for this overestimation. You should use the adjusted *R*-square as a measure of variance accounted for rather than the unadjusted *R*-square. You also should pay attention to the standard error. The standard error gives you an indication of how much variability there is around the calculated regression line. The lower the value, the better it is.

Regression Weights Regression weights are used to interpret the results from a multiple regression analysis. There are two types of regression weights: raw and standardized. A raw regression weight (*b*) is calculated based on the raw scores entered into the regression analysis. A standardized regression weight is calculated after your raw scores have been transformed to standard scores. The standardized regression weights are known as **beta weights** (abbreviated with the Greek symbol β). When you use a computer program (such as SPSS for Windows) to conduct a regression analysis, a *t* value for each regression weight should be provided. The *t* value tells you whether the regression weight is statistically significant.

For most applications in psychological research, you should use the standardized regression weights (beta weights) because they can be directly compared even if the variables to which they apply were measured on very different scales. For example, the beta weights given to variables such as intelligence, GPA, and socioeconomic status (which are all measured on different, nonequivalent scales) can be directly compared whereas the *b* weights cannot. Only when your variables are measured on the same standard scale should you use the raw score regression weights.

Interpretation of Regression Weights If your regression analysis is significant, you may want to know how much of the variability in the criterion variable can be accounted for by variation in each predictor. Avoid using the beta weights for this. A beta weight is not an index of the *unique* contribution of a given predictor to variability in the dependent variable.

A beta weight for a given predictor variable may be high because either the predictor *directly* produces most of the variance in the dependent variable or it is merely *correlated* with another, effective predictor variable (Tabachnick & Fidell, 2001). Similarly, a beta weight for a given predictor variable may be low, and yet the predictor may have a strong causal influence on the dependent variable. This situation can occur when other predictor variables in the analysis correlate with the effective variable. The analysis may then mistakenly assign weight to the correlated variables instead of to the effective one. In such cases, the correlated variables are termed *suppressor variables* because they mask (or suppress) the effect of the effective variable.

An alternative to using beta weights to determine the unique contribution of each predictor is the *squared semipartial correlation* (Tabachnick & Fidell, 2001). Simply square the part correlation for each variable to obtain the squared semipartial correlation. These numbers represent the amount of variability accounted for by each variable.

Squared semipartial correlations need not sum to *R*-square. If the sum of semipartial correlations is less than *R*-square, then the difference between the two numbers represents the shared variance (Tabachnick & Fidell, 2001). In some cases, the sum of squared semipartial correlations can be larger than *R*-square.

An Example of Multiple Regression Because multiple regression is commonly used and underlies many other multivariate statistics, we shall present an extended example. Recall the previously mentioned study of the factors correlating with a student's attitude toward seeking counseling for personal problems. David Vogel and Stephen Wester (2003) actually conducted such a study. They had 209 college students (143 male and 66 female) complete a measure of their attitude toward seeking counseling. The students also completed measures concerning their willingness to disclose personal information about themselves, how risky they felt such disclosures were, and whether they had sought counseling in the past. These latter measures, along with participant gender, yielded five predictor variables for a multiple regression analysis: willingness to self-disclose personal information, risk of disclosing emotional information, utility of disclosing emotional information, gender, and previous counseling.

The results showed a significant regression analysis, $F(5, 190) = 25$, $p < .001$. The R^2 was .40, and the adjusted R^2 was .39, indicating that 39% of the variance was accounted for by the regression analysis. Table 15-2 shows how each predictor variable related to the dependent variable and whether each predictor variable contributed significantly to the regression analysis. The first column shows the raw score regression weights. Remember, we don't use them to interpret a regression analysis. The third column shows the standard error of the regression weights. The beta weights (standardized regression weights) are shown in the fourth column. These are the ones that we use to interpret the regression analysis.

As you can see, the strongest predictor of attitude toward seeking counseling is a person's willingness to disclose distressing information. The positive regression weight tells us that individuals who are more willing to self-disclose are more likely to have a positive attitude toward seeking counseling. Another strong predictor is the anticipated utility of self-disclosing emotional information ($\beta = .24$), indicating that individuals who see emotional self-disclosure as useful have a more positive attitude toward seeking help. Notice that the anticipated risk of emotional

TABLE 15-2 Results From Vogel and Wester's (2003) Multiple Regression Analysis

CRITERION VARIABLE	b	SE_b	β	t	p
Self-disclosure of distressing information	4.1	.84	.29	4.8	<.001
Anticipated risk of emotional self-disclosure	−1.9	.63	−.18	−3.1	<.01
Anticipated utility of emotional self-disclosure	2.9	.69	.24	4.1	<.001
Participant gender	7.5	1.7	.27	4.5	<.001
Previous counseling	−5.4	1.6	−.2	−3.4	<.01

SOURCE: Vogel and Wester, 2003; reprinted with permission.

self-disclosure is negatively related to help-seeking attitude ($\beta = -.18$), indicating that those who see emotional self-disclosure as risky have a more negative attitude toward seeking help.

QUESTIONS TO PONDER

1. For what research applications would you use the various types of multiple regression analysis?
2. How are multiple R, R^2, and adjusted R^2 used to interpret the results from a multiple regression analysis?
3. What is the difference between the raw and standardized regression weights, and why are the standardized weights used when interpreting the results from a regression analysis?
4. What is the squared semipartial correlation, and when is it used?

Discriminant Analysis

Discriminant analysis is a special case of multiple regression. It is used when your dependent variable is categorical (e.g., male–female or Democrat–Republican–Independent) and you have several predictor variables. Discriminant analysis allows you to predict membership in a group (one of the discrete categories of your dependent variable) based on knowledge of a set of predictor variables. You can use discriminant analysis to identify a simple rule for classifying participants into groups or to determine which of your predictor variables contributes most heavily to the separation of groups. The analysis works by forming *discriminant functions*. For each dependent variable group, a discriminant function score is calculated according to the following formula (Tabachnick & Fidell, 2001):

$$D_{\mathrm{i}} = d_{i1}z_1 + d_{i2}z_2 + \cdots + d_{in}z_n$$

where D_i is the discriminant function score calculated for each participant, d_i is the regression weight, and z is the standardized raw score on a particular predictor. In discriminant analysis, a new variable (D_i) is calculated for each participant. This variable is the best linear combination of predictor variables, just as in multiple regression. When the discriminant function scores have been calculated for each group, a centroid can then be determined. The *centroid* is simply the average of the discriminant function scores within a group.

More than one discriminant function can link your predictors with your dependent variable. However, the number of functions is limited to the number of predictors or to the number of levels of the dependent variable minus 1, whichever is smaller. For example, if you had seven predictors and three levels of the dependent variable, the number of possible functions is 2 (or 3 − 1). Each discriminant function represents a different linkage between the predictors and the dependent variable. The first one calculated maximizes the separation between levels of the dependent variable. Subsequent functions represent progressively weaker linkages between the predictors and the dependent variable.

Because the computations needed to perform a discriminant analysis are complex, you will probably use a computer program to conduct a discriminant analysis. SPSS conducts a discriminant analysis within its Analyze subprogram. The output of the SPSS analysis gives you several important pieces of information. First, the output will indicate the number of discriminant functions extracted, along with tests of statistical significance. Second, you can request several other statistics needed to interpret your results. These include the standardized discriminant function coefficients (analogous to beta weights) and pooled within-groups correlations between the discriminant functions and predictor variables (structure correlations).

You can use a discriminant analysis in two ways. First, you can evaluate the amount of variability accounted for by each function. You would do this by conducting a dimension reduction analysis that provides a canonical correlation coefficient and significance tests for each function. The squared canonical correlation coefficient gives a measure of the amount of variability accounted for by a specific function. By looking at the dimension reduction analysis, you can determine the significance of each function and the amount of variability accounted for by each function.

The second way you can use the discriminant analysis is to evaluate the degree of contribution of each predictor (within a function) to the separation of groups. One strategy is to look at the standardized discriminant function coefficients. However, these weights (like beta weights) do not reveal how much each individual predictor contributes to variation in the dependent variable. Another strategy is to look at the structure correlations, which can be interpreted much like factor loadings. By convention, you typically consider those structure correlations that exceed .30. The structure correlations can help you determine what each discriminant function represents. However, they are not good indicators of the predictor's degree of unique contribution to discriminating among dependent variable groups (Tabachnick & Fidell, 2001).

Rather than looking at beta weights or structure correlations, you could conduct a set of specific contrasts in which each dependent variable group is contrasted with all others. You then look for which predictor variables separate a particular group from the rest (Tabachnick & Fidell, 2001). This procedure is too complex to fully describe here. (See Tabachnick & Fidell, 2001.)

Canonical Correlation

Multiple regression determines the relationship between a set of variables (predictors) and a *single* dependent variable. To determine the relationship between a set of predictors and a *set* of dependent variables, you use canonical correlation. Canonical correlation works by creating two new variables for each subject, called *canonical variates*. A canonical variate is computed both for the dependent and predictor sets. The canonical variate is simply the score predicted from a regression equation based on the variables within a set. The correlation between the two canonical variates is the *canonical correlation*.

Canonical correlation does not appear much in published psychological literature because, at this point in its development, it is a purely descriptive strategy (Tabachnick & Fidell, 2001). It can be used to describe the relationship between two sets of variables, but it cannot be used to infer causal relationships. Consequently, this technique is not discussed further. If you want to know more about the technique, see Tabachnick and Fidell (2001) and Levine (1977).

QUESTIONS TO PONDER

1. What is discriminant analysis, and when is it used?
2. What do discriminant functions tell you?
3. What are the two main applications of discriminant analysis?
4. What is canonical correlation, and when is it used?

EXPERIMENTAL MULTIVARIATE STATISTICAL TESTS

In this section we review two multivariate statistical analysis techniques used to analyze data from multivariate experimental designs: multivariate analysis of variance and multiway frequency analysis.

Multivariate Analysis of Variance

Assume that you are required to conduct an experiment for a senior thesis. Your major area of interest is in the development of a concept of death among school-aged children. You have reviewed the literature and have found most of the current research to be correlational. You decide there is room for some experimental work in the area, but you also decide to draw on the existing correlational research to help you develop your measures. You find that the previous research suggests that several important measures should be applied to assessing children's concepts of death. So you decide to include three measures in your experiment.

The existing literature suggests that a child's concept of death can be accelerated by exposure to experience with the concept of death. So you decide to conduct a single-factor experiment with three groups. The first group is simply exposed to a film about a character who dies. The second group role-plays a dying animal. The third group, a control group, receives no special treatment.

After running your experiment, you are faced with the problem of how to analyze your three dependent measures. Of course, you could simply conduct three separate one-factor ANOVAs. You are uncomfortable with this strategy because the existing literature indicates that your three chosen measures are correlated. You might miss some important relationships among your variables if you simply use a series of univariate tests. In this situation, a viable alternative is to use a MANOVA to analyze your data. Like canonical correlation and discriminant analysis, MANOVA operates by forming a new linear combination of dependent variables for each effect in your design. For example, for a two-factor between-subjects design, a different linear combination of scores is formed for each of the two main effects and for the interaction.

An Example of MANOVA Suppose you conducted the one-factor experiment looking at the effect of a training program on children's concepts of death. Your measure of the concept of death consisted of a questionnaire containing several important questions concerning death (e.g., "What happens when you die?" "What can you do to bring something dead back to life?" and "Do dead people feel pain?"). Children simply answered the questions. Independent raters then indicated how mature the concept of death was in each response. Because your measures were related to the same concept, you decide to use a MANOVA rather than separate ANOVAs to analyze the data.

Table 15-3 shows some hypothetical data that might be generated from such a study. Table 15-4 shows part of the output from an SPSS MANOVA analysis of these data. The top part of Table 15-4 shows the multivariate tests of significance. Although the results from a number of such tests are shown, you decide to use the Wilks's test (the reasons why you choose one test over another are not important here because, in most cases, there will be little difference among them). The Wilks's test indicates that the effect of the treatment was significant, $F(6, 20) = 13.14$, $p < .001$. This tells you that your independent variable reliably affected the value of the linear dependent variable created in the MANOVA. When you conduct a single-factor MANOVA, you get essentially the same analysis as a canonical correlation analysis. As in canonical correlation, SPSS MANOVA extracts as many discriminant functions as possible (two in this case). The second section of Table 15-4 shows the statistics relevant to the discriminant functions extracted. Here, the canonical correlations are presented ("Canon Cor.") along with the percentage of variance accounted for ("Pct.") and the associated eigenvalues. Eigenvalues are not discussed here. See Tabachnick and Fidell (2001) for a discussion of these values.

TABLE 15-3 Data From Hypothetical Experiment on Developing a Concept of Death

	FILM			ROLE PLAYING			CONTROL		
Subject	X_1	X_2	X_3	X_1	X_2	X_3	X_1	X_2	X_3
1	6	3	5	8	7	9	1	2	1
2	5	5	3	7	9	5	2	2	4
3	6	4	2	9	9	7	1	1	2
4	4	2	2	6	8	8	3	3	3
5	3	2	6	7	8	9	4	1	2

NOTE: X_1, X_2, and X_3 refer to the three dependent measures listed in the text.

TABLE 15-4 Partial SPSS Output for Hypothetical Experiment on Developing a Concept of Death

MULTIVARIATE TESTS OF SIGNIFICANCE (S = 2, M = 0, N = 4)

Test Name	*Value*	*Approx.*	*Hypoth.* df	*Error* df	*Sig. of* df
Pillai's	1.21442	5.66832	6.00	22.00	.001
Hotelling's	17.82643	26.73964	6.00	18.00	.000
Wilks's	.03962	13.41251	6.00	20.00	.000
Roy's	.94583				

Note: *F* statistic for Wilks's Lambda is exact.

EIGENVALUES AND CANONICAL CORRELATIONS

Root No.	*Eigenvalue*	*Pct.*	*Cum. Pct.*	*Canon Cor.*
1	17.459	97.940	97.940	.973
2	.367	2.060	100.000	.518

DIMENSION REDUCTION ANALYSIS

Roots	*Wilks's* L	F *Hypoth.*	df	*Error* df	*Sig. of* F
1 TO 2	.03962	13.41251	6.00	20.00	.000
2 TO 2	.73140	2.01981	2.00	11.00	.179

ROY–BARGMAN STEPDOWN *F* TESTS

Variable	*Hypoth.* MS	*Error* MS	*Stepdown*	F *Hypoth.* df	*Error* df	*Sig.* *of* F
DIE	33.80000	1.56667	21.57447	2	12	.000
LIFE	13.73684	1.06132	12.94322	2	11	.001
FEEL	7.89738	2.47695	3.18835	2	10	.085

CORRELATIONS BETWEEN DEPENDENT AND CANONICAL VARIABLES

	Canonical Variable	
VARIABLE	1	2
DIE	−.436	.870
LIFE	−.735	−.198
FEEL	−.376	−.025

The third section of Table 15-4 presents the results of a dimension reduction analysis. This analysis reveals that only the first function was significant ($p < .001$).

The significant *F* ratio in the multivariate test indicates a reliable effect of the training program on concepts of death. As a next step, you assess each dependent variable's contribution to this significant effect. There are several ways to do this. You could simply look at the univariate *F* tests produced by SPSS. This strategy has limitations, especially if your dependent variables are correlated with one another (Tabachnick & Fidell, 2001). An alternative strategy is to examine the *Roy–Bargman stepdown analysis* shown in the fourth section of Table 15-4.

In the Roy–Bargman stepdown analysis, the first dependent variable is entered and tested for significance. Then the second variable is entered, and the first variable is treated as a covariate. This test tells you whether each dependent variable explains variability over and above the variability already explained by those previously entered (Tabachnick & Fidell, 2001). The Roy–Bargman stepdown analysis is similar in concept to hierarchical regression. The main drawback to this analysis is that it can be used only when you can specify the order in which variables are entered. In the absence of a theoretically or empirically based order of entry, the Roy–Bargman test should not be used.

We should note that you cannot obtain the Roy–Bargman stepdown analysis directly from SPSS's pull-down menu system. Instead, you will have to use the syntax editor and enter a series of commands to obtain this analysis. To do this, you will need to know the appropriate commands and how to structure them. Instructions on how to execute these commands can be found within SPSS for Windows.

Finally, you also could look at the structure correlations shown at the bottom of Table 15-4 (labeled "Correlations Between Dependent and Canonical Variables"). The structure correlations are similar to factor loadings and can be interpreted as such.

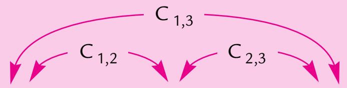
In addition to these analyses, you might want to conduct some post hoc analyses to determine the specific effect of the independent variable on the dependent variables. These tests are similar to those used in univariate ANOVA but are more complex.

Using MANOVA for Within-Subjects Designs Chapters 10 and 14 discussed within-subjects designs and analyses. Each subject in a within-subjects design is exposed to all levels of your independent variable. Chapter 14 indicated that a repeated-measures ANOVA can be used to analyze data from this design. Data from within-subjects (and mixed) designs also can be analyzed with MANOVA.

The within-subjects ANOVA assumes homogeneity of both within-cell variances and within-cell covariances. The first assumption states that variance should be homogeneous across treatments. This assumption is common between the between-subjects and within-subjects analyses. However, the assumption of homogeneity of covariance is special to the within-subjects ANOVA. The following example illustrates the idea of homogeneity of covariances.

Figure 15-4 shows a simple one-factor within-subjects design. Notice that the behavior of each subject is evaluated under each level of the independent variable.

FIGURE 15-4 Three-treatment within-subjects design showing covariances.



Subject	Level 1	Level 2	Level 3
1	$X_{1,1}$	$X_{2,1}$	$X_{3,1}$
2	$X_{1,2}$	$X_{2,2}$	$X_{3,2}$
3	$X_{1,3}$	$X_{2,3}$	$X_{3,3}$
4	$X_{1,4}$	$X_{2,4}$	$X_{3,4}$
5	$X_{1,5}$	$X_{2,5}$	$X_{3,5}$

Note: $X_{i,j}$ represents a subject's score, where *i* = the level of the independent variable and *j* = the subject number. $C_{i,j}$ indicates the covariances that could be calculated, where *i* and *j* are the levels of the independent variable.

This being the case, you can form pairs of conditions and subtract the scores associated with each subject within those two conditions (Keppel, 1982). You can then obtain variances based on the resulting difference scores. This variance is called *covariance*. In the design illustrated in Figure 15-4, you can form three pairs of conditions (C_1 with C_2, C_2 with C_3, and C_1 with C_3) and three concomitant covariances. If the covariances are not homogeneous, then the homogeneity-of-covariance assumption has been violated.

In a between-subjects ANOVA, mild to moderate violations of the homogeneity assumption do not significantly affect the validity of the statistical test. In a within-subjects ANOVA, however, violations of the homogeneity assumption lead to a serious positive bias: rejecting the null hypothesis more often than you should at a given alpha level. Thus, you are making more Type I errors than are acceptable.

Another problem with the univariate within-subjects analysis is that when you add a second independent variable, the analysis becomes somewhat controversial. The controversy surrounds selecting an error term appropriate to test the main effects and interactions. You can select a "pooled" error term and use this same term for all the within-subjects factors, or you can use separate error terms for each. Unfortunately, the two choices may lead to different outcomes of the statistical analysis, and there is little agreement on which choice is best.

Because MANOVA circumvents the problems with the homogeneity assumptions and error-term selection, it has been suggested as an alternative to standard within-subjects ANOVA (O'Brien & Kaiser, 1985; Tabachnick & Fidell, 2001). In the MANOVA, the repeated measures taken on each subject are treated as correlated dependent variables and analyzed accordingly. Rather than assuming homogeneity of covariance, MANOVA takes into account the covariances actually present in the data. For information on using MANOVA to analyze within-subjects designs, see O'Brien and Kaiser (1985).

QUESTIONS TO PONDER

1. When would you use a multivariate analysis of variance (MANOVA) to analyze your data?
2. How are the results of a MANOVA interpreted?
3. Why would you use a MANOVA to analyze data from a within-subjects design?

Multiway Frequency Analysis

Most of the powerful inferential statistics discussed in this chapter and in Chapter 14 require that your variables be measured along at least an interval scale. However, there are research situations in which you may want to measure or manipulate categorical variables (e.g., sex of subject). Statistics such as ANOVA, MANOVA, and multiple regression are not appropriate to analyze such data. For these cases, **multiway frequency analysis** is an alternative. A specific type of multiway frequency analysis used for categorical or qualitative variables is loglinear analysis.

Loglinear analysis is analogous to chi-square (see Chapter 14) in that you use observed and expected frequencies to evaluate the statistical significance of your data. An important difference between chi-square and loglinear analysis is that loglinear analysis can be easily applied to experimental research designs that include more than two independent variables whereas chi-square is normally limited to the two-variable case.

Applications of Loglinear Analysis Loglinear analysis has a wide range of applications. You can use loglinear analysis if you conducted a correlational study with several categorical variables; loglinear analysis is well suited to this task. You also can use loglinear analysis if you conducted an experiment including a categorical dependent variable (e.g., guilty/not guilty) even if your independent variables were quantitative.

Loglinear analysis is also a useful tool for testing and building theoretical models. In this application, you specify how variables should be entered into the analysis for the models that you wish to test. Loglinear analysis is then used to test the relative adequacy of each model.

Finally, because loglinear analysis is a nonparametric statistic, you can use it if your data violate the assumptions of parametric statistics (such as ANOVA). In this instance, you can use loglinear analysis even if your dependent variable was measured along an interval or ordinal scale (Tabachnick & Fidell, 2001). However, one important requirement must still be met.

Like chi-square, loglinear analysis uses observed and expected cell frequencies to compute your test statistic. To obtain valid results, your expected cell frequencies must be relatively large. Tabachnick and Fidell (2001) recommend having five times as many subjects as cells to ensure adequate expected frequencies. So, for example, if you have a $2 \times 2 \times 2$ design, you should have $2 \times 2 \times 2 \times 5 = 40$ subjects to ensure sufficiently large expected cell frequencies.

You should inspect all expected frequencies to ensure that they are all larger than one and that no more than 20% of them fall below five (Tabachnick & Fidell, 2001). If you find small expected frequencies, Tabachnick and Fidell suggest several remedies. First, you could accept and live with the reduced power of the analysis caused by low expected frequencies. Second, categories can be collapsed or combined to increase frequencies within each category. Third, you could delete variables to reduce the number of categories in your analysis. This latter strategy should be done with caution, and you should not delete variables that are correlated with other variables in the analysis (Tabachnick & Fidell, 2001).

How Loglinear Analysis Works When you use ANOVA or multiple regression to analyze your data, your analysis uses group means as a basis for analysis. When you have categorical data, however, you must deal with proportions instead of means. For example, if you used a three-factor design with a categorical dependent variable (yes/no), you would summarize your data according to the proportion of subjects falling into each category.

In a standard analysis of proportional data from a two-factor experiment in which you are interested in a single relationship between two variables, chi-square is used to evaluate that relationship. However, when evaluating more than one relationship (i.e., two main effects and an interaction), chi-square is not the best test statistic because the component chi-squares do not sum to the total chi-square (Tabachnick & Fidell, 2001). In this case, a likelihood ratio (G^2) is used in place of chi-square. Although similar to chi-square (in that cell expected and observed frequencies are compared), G^2 involves taking the natural log (ln) of the ratio of observed cell frequency to the expected cell frequency, according to the following formula (Tabachnick & Fidell, 2001):

$$G^2 = 2(f_0)\ln(f_o/f_e)$$

where f_o is the observed cell frequency and f_e is the expected cell frequency. A G^2 is computed for each main effect and interaction in your design and is interpreted in the same way as chi-square (using the chi-square tables to establish statistical significance).

Space limitations here preclude a detailed description of all the applications of loglinear analysis or how loglinear analysis is used. If you need to use loglinear analysis, detailed discussions can be found in Agresti and Finlay (1986) and Tabachnick and Fidell (2001).

QUESTIONS TO PONDER

1. What is multiway frequency analysis, and when is it used?
2. What is loglinear analysis and when is it used?
3. What is G^2 and how is it used in loglinear analysis?

MULTIVARIATE STATISTICAL TECHNIQUES AND CAUSAL MODELING

In this section we discuss two applications of multivariate statistics to causal modeling: path analysis and structural equation modeling.

Path Analysis

Path analysis applies multiple regression techniques to causal modeling. For example, suppose you are interested in determining how attitudes and behaviors relate to one another. You have reviewed the literature and have come across the "theory of reasoned action" by Fishbein and Ajzen (1975). This theory postulates that attitudes are evaluative dimensions that, along with subjective norms (such as knowing what your friends are going to do), mediate between other variables (such as sex) and behavioral intentions (i.e., what you specifically intend to do). According to the theory, behavioral intentions in turn determine behavior.

You decide to test the limits of the Fishbein and Ajzen (1975) theory by measuring each of the important components of the theory with a questionnaire. The topic that you have chosen is the relationship between attitudes toward college and actual college attendance. To test the theory, you collect data on attitudes toward attending college, subjective norms about college, a specific intention to attend college, and actual college-attendance behavior.

After you have collected your data, you now face the task of analyzing them. Whereas you could simply compute bivariate correlation coefficients between all variables, the disadvantages of this approach have been discussed. Besides, simple correlational analyses will not allow you to evaluate possible causal relationships among your variables. As an alternative, consider using path analysis.

Unlike the other analytic techniques already discussed, path analysis is not a statistical procedure in and of itself. Rather, it is an application of multiple regression techniques to the testing of causal models. Path analysis allows you to test a model specifying the causal links among variables by applying simple multiple regression techniques.

Always remember that path analysis is designed to test causal models, not to sift through data for interesting relationships among variables. Developing a clearly articulated causal model is crucial in path analysis. The model should not rest on flimsy ideas and unsupported conjecture. Instead, the causal relationships proposed in the model should rest on a strong theoretical or empirical base.

Translating theoretical propositions into a clearly defined path model can be tricky. You always are tempted to determine how to measure your variables first and then derive the model. This method may not be the best. It may limit the possible causal relationships within your model and consequently may not allow you to adequately test your theory. Instead, first develop a list of the causal links among variables as suggested by your theory (Hunter & Gerbing, 1982). Then show these links among variables in a *path diagram*. After developing the path model and diagram, you can then decide how to measure your variables.

Causal Relationships The heart of path analysis is developing a causal model and identifying causal relationships. Causal relationships among variables can take many forms. The simplest of these is shown in panel (a) of Figure 15-5, where Variable A (independent variable) causes changes in Variable B (dependent variable). Another possible causal relationship is shown in panel (b). Here, two variables impinge on Variable B. This model suggests that variation in the dependent variable has multiple causes. These causal variables can be uncorrelated as shown in panel (b). Panel (c) shows a situation in which two variables believed to cause changes in the dependent variable are correlated. In Figure 15-5 (and in path analysis, in general), straight arrows denote causal relationships and are called *paths*. Curved, double-headed arrows denote correlational relationships.

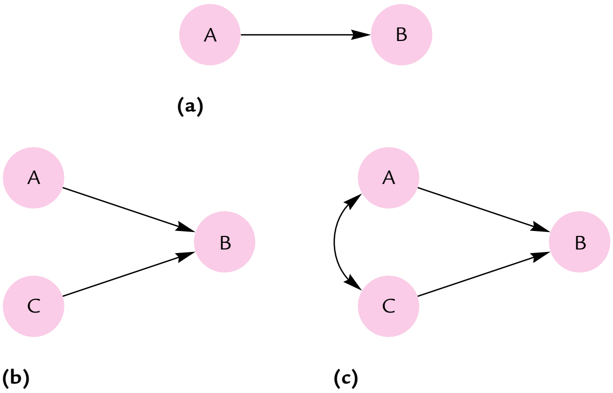
The simple causal relationships just described can be combined to form more complex causal models. One such model is the *causal chain* in which a sequence of events leads ultimately to variation in the dependent variable. To illustrate a simple causal chain, consider a modification of a previous example in which you were trying to determine what variables correlated with SAT scores.

Suppose you believe that parental education (PE) and student motivation (SM) relate to variation in SAT scores. You have reason to believe that a causal relationship exists. So you develop a causal model like the one illustrated in panel (a) of Figure 15-6. Your model suggests that PE causes changes in SM, which then causes changes in SAT scores. Notice that you are proposing that PE does not directly cause changes in SAT but rather operates through SM.

When developing simple causal chains (and more complex causal models), keep in mind that the validity of your causal model depends on how well you have done your homework and conceptualized your model. Perhaps SM does not directly cause changes in SAT scores as conjectured but rather operates through yet another variable, such as working hard (WH) in class. Panel (b) of Figure 15-6 shows a causal chain including WH. If you excluded WH from your model, the causal relationships and the model you develop may not be valid.

You can progress from simple causal chains to more complex models quite easily. Figure 15-7 shows three examples of more complex causal models. In panel (a), the causal model suggests Variables A and B are correlated (indicated with the curved

FIGURE 15-5 Three possible causal relationships. (a) Variable A causes changes in B; (b) uncorrelated Variables A and C contribute to changes in the value of B; (c) correlated Variables A and C cause changes in the value of B.



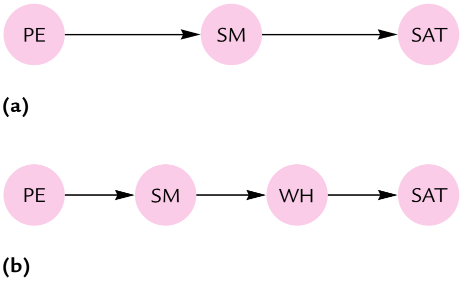


FIGURE 15-6 (a) Three-variable causal chain and (b) four-variable causal chain.

arrow). Variable A is believed to exert a causal influence on Variable C, and B on D. Variable D is hypothesized to cause changes in C, and both D and C are believed to cause changes in E.

Types of Variables and Causal Models Variables A and B in panel (a) of Figure 15-7 are called *exogenous variables*. Exogenous variables begin the causal sequence. Notice that no causal paths lead to Variable A or B. All the other variables in the model shown in panel (a) are *endogenous variables*. These variables are internal to the model,

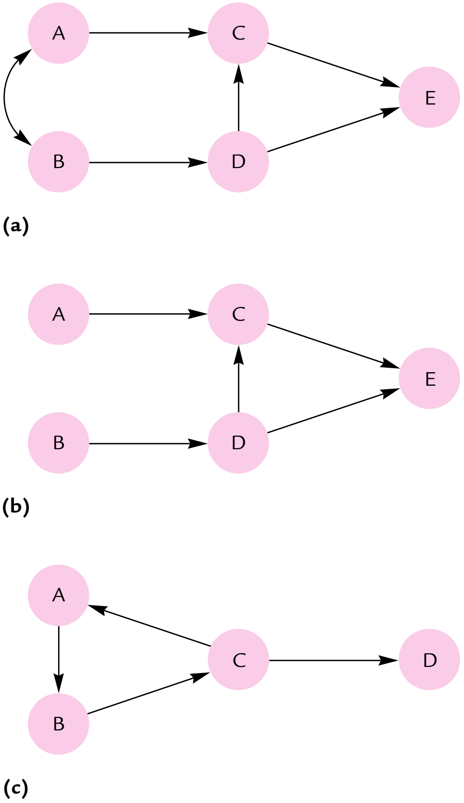


FIGURE 15-7 Three complex causal models.

and changes in them are believed to be caused by other variables. Variables C, D, and E are all endogenous variables. Panel (b) of Figure 15-7 shows essentially the same model as panel (a), except that the two exogenous variables are not correlated in panel (b).

The models in panels (a) and (b) are both known as *recursive models*. Notice that there are no loops of variables. That is, causal relationships run in only one direction (e.g., D causes C, but C does not cause D). In contrast, panel (c) of Figure 15-7 shows a *nonrecursive model,* which has a causal loop. In this case, Variable A is believed to be a cause of C (operating through B), but C also can cause A. In general, recursive models are much easier to deal with conceptually and statistically (Asher, 1976).

Estimating the Degree of Causality After you have developed your causal models and measured your variables, you then obtain estimates of the causal relationships among your variables. These estimates are called *path coefficients*. Figure 15-8 shows a causal model with the path coefficients indicated for each causal path.

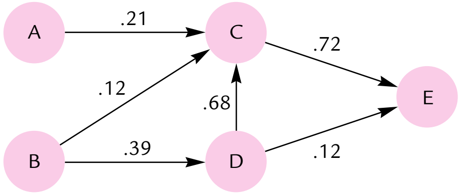
Path coefficients are determined by using a series of multiple regression analyses. Each endogenous variable is used as a dependent variable in the regression analysis. All the variables in the model that are assumed to impinge on the dependent variable are used as predictors. For example, the path coefficients for A–C and D–C in Figure 15-8 are obtained by using C as the dependent variable and A, B, and D as predictors. The path coefficients are the standardized regression weights (beta weights) from these analyses.

Interpreting Path Analysis Path analysis is used to test the validity of a presumed causal model. To that end, you look at the path coefficients and determine whether the pattern expected by the model has emerged. In addition to looking at the path coefficients (which give you estimates of the direct effects of variables on other variables), you also decompose the paths into indirect effects. Decomposition can be done according to Wright's rules (see Asher, 1976, for how this can be done).

QUESTIONS TO PONDER

1. What is path analysis, and how is it used in the research process?
2. Why is it important to develop a causal model when using path analysis?
3. What are the different types of variables used in a path analysis?
4. How are path coefficients used to interpret a path analysis?

FIGURE 15-8 Path diagram showing path coefficients.



Structural Equation Modeling

Structural equation modeling (SEM) is a variant of path analysis. With path analysis, variables that are directly observed and measured are included in the analysis (Streiner, 2006). Sometimes, however, you deal with constructs that are not directly observable but rather are manifested in a number of behaviors (Streiner, 2006). For example, depression is a hypothetical construct that is not directly observable. Instead, we can measure several behaviors that relate to depression (e.g., suicidal thoughts, loss of energy, and sleep disturbances). One advantage of SEM over path analysis is that it allows you to evaluate hypothetical constructs within the models that you test (Streiner, 2006). In the language of SEM, those variables in a model that are not directly observable or measurable are called **latent variables**.

SEM is normally used as a confirmatory procedure and not an exploratory one (Garson, 2006). G. David Garson suggests that there are three confirmatory applications of SEM:

1. *Strictly confirmatory approach:* You test a model to see if data that you collected are consistent with the predictions of the model.
2. *Alternative models approach:* You test two or more alternative models to see which one (if any) best fits the data collected.
3. *Model development approach:* You use SEM to develop a model by combining exploratory and confirmatory approaches. You then test the model, and if you find that it does not fit the data very well, you make modifications to the model and retest it. You keep doing this until you find the model that best fits the data.

An important fact to keep in mind about SEM is that it requires you, as the researcher, to develop a model to test based on existing theory and research. You do not simply go on a "fishing expedition" by throwing variables into an SEM analysis and hope to find meaningful relationships and causal connections. However, as noted above, SEM can serve an exploratory function. Even this, however, requires that you specify a coherent model to be tested.

Developing a model for SEM analysis starts with a verbal statement of how variables relate (e.g., according to a theory) (Hershberger, Marcoulides, & Parramore, 2003). Next, you draw out the model using boxes, ovals, and arrows. Boxes represent variables that you will measure or have measured; their names are written inside the boxes. Circles denote latent variables. Arrows specify the relationships among variables: straight arrows for causal relationships and curved arrows for correlational relationships (Hershberger et al., 2003).

As you saw in our discussion of path analysis, path coefficients are derived by using multiple regression analysis. In SEM specialized statistical techniques are used to derive coefficients. The most popular program used for this purpose is LISREL, published by Scientific Software International (SSI). The LISREL package includes programs for various types of modeling. It is beyond the scope of this chapter to explore LISREL in detail. (For details, visit the SSI Web site at http://www.ssicentral.com.)

MULTIVARIATE ANALYSIS: A CAUTIONARY NOTE

The bare-bones overview of the major multivariate techniques in this chapter could not provide a detailed discussion of the controversies surrounding the use of these tests. Although some researchers characterize multivariate statistics such as MANOVA as "a powerful and rich methodology to characterize group difference" (Bray & Maxwell, 1982 p. 363), others advocate extreme caution in applying multivariate statistics (Hunter, 1987). In our experience, caution is called for when using multivariate statistics.

The computer can analyze multivariate data quickly and efficiently. In fact, the computer has made multivariate statistical techniques readily available to most researchers. With such easy access inevitably comes misapplication. The computer can grind out pages and pages of output in an amazingly short period of time. It cannot interpret the results for you, however. Much of the controversy over the use of multivariate statistics lies in the area of interpretation.

In multiple regression, for example, is it really better to use the standardized regression weights or the unstandardized weights to interpret the data? Or should you heed the advice of Tabachnick and Fidell (2001) to calculate semipartial correlations? In discriminant analysis, should you use the standardized coefficients or the structure correlations? In factor analysis, should you use orthogonal or oblique rotation? Unfortunately, there are no universally agreed-on answers to these and other major questions concerning multivariate analyses.

This chapter has discussed some of the advantages of multivariate statistics over univariate statistics. Unfortunately, multivariate statistics are not simple substitutions for univariate statistics. In a univariate ANOVA, for example, the effect of an independent variable on a dependent variable is evaluated by determining if the observed means change as a function of changes in the independent variable. Fine-grained analyses are then conducted to determine which means differ significantly. MANOVA, however, produces results that are more difficult to interpret. Instead of performing simple fine-grained analyses to localize significant effects, you set up contrasts with discriminant analyses. In short, interpreting results from a MANOVA is more complex than interpreting results from univariate ANOVAs.

The most prudent thing that you can do at this point is to spend some time learning about the intricacies of these tests so that you can identify and avoid these hidden traps. Thoroughly familiarize yourself with the assumptions of multivariate statistics and with how these statistics operate before attempting to use them.

Also, pay close attention to how you design your study. In many cases, failure to uncover relationships with a multivariate test is caused by faulty logic more during the design phase than during the analysis phase (Asher, 1976). Multivariate statistics cannot make sense out of poorly conceptualized and measured variables. There is no substitute for a carefully designed multivariate study that has a sound theoretical or empirical base and a well-defined measurement model.

QUESTIONS TO PONDER

1. What is structural equation modeling, and how does it differ from path analysis?
2. What are the three confirmatory applications of structural equation modeling?
3. What is a latent variable, and how is it used in structural equation modeling?
4. Why should you exercise caution when using multivariate data analysis?

SUMMARY

Whenever you include several related measures in the same study, you are using a multivariate design. Analysis of your data is then done with one of the many multivariate statistical tests. There are multivariate tests for experiments with multiple dependent measures (MANOVA and MANCOVA) and tests for correlational designs (multiple regression, canonical correlation, discriminant analysis, and factor analysis). These tests allow you to identify complex relationships while controlling statistical errors.

Like univariate and bivariate statistics, multivariate statistics make assumptions that must be met. Your variables must be linearly related and normally distributed. Violations of these assumptions may lead to invalid conclusions. In addition, you must identify and deal with both univariate and multivariate outliers because outliers can drastically affect the correlations used to compute multivariate statistics. Outliers can be handled either by deleting cases or variables or by using an appropriate data transformation. Also, make an effort to develop a sound measurement model to avoid the problem of error of measurement. Error of measurement can lead to an unacceptable number of Type II errors.

Factor analysis is used either to reduce a large set of variables to a smaller set or to confirm that certain variables measure the same underlying factor. Two types of factor analysis are principal components and principal factors. The difference between the two is found in the values placed on the diagonal of the correlation matrix used to extract factors. In principal components analysis, 1s are placed on the diagonal. In principal factors analysis, communalities are placed on the diagonal. These communalities are difficult to derive before factor analysis. Principal components analysis is therefore more popular.

Factor analysis extracts as many significant factors as possible. For each factor, a variable will have a certain loading. These loadings are the correlations between the original variable and the factor extracted. You should rotate the factors before interpreting the loadings. The most popular rotation method is varimax.

Partial and part correlation analyses are used when you want to evaluate the relationship between two variables while controlling for a third. Partial correlation evaluates the relationship between two variables with the effect of a third variable removed from both of the variables being correlated. A part correlation (also known as a semipartial correlation) evaluates the relationship between two variables with the effect of a third removed from only one of them.

Multiple regression is used when you have identified a single, continuously measured dependent variable and several predictor variables. The analysis operates by determining regression weights based on the correlations among your variables. Two types of regression weights are raw and standardized weights. In general, you should use the standardized weights to help interpret your results. Unfortunately, the standardized weights do not tell you how much each variable contributes to explaining variability in your dependent variable. To do this, you should calculate the squared semipartial correlations. These correlations can then be used to determine the degree to which each variable independently contributes to variation in the dependent variable.

Discriminant analysis is an extension of multiple regression. It is used when your dependent variable is categorical. Essentially, you are trying to predict group membership based on knowledge of your predictor variables. Interpretation is made by examining the amount of variability accounted for by each of the extracted discriminant functions and by setting up contrasts between your dependent variable groups. The former tells you how important each function is, and the latter how important each variable is to the solution.

In situations in which you have two sets of variables, the analysis of choice is canonical correlation. One set of variables may be identified as the dependent variable set and the other, the predictor variable set. Canonical correlation computes for each set a new variable called a canonical variate. The correlation between the canonical variates is the canonical correlation. Interpretation of a canonical analysis can be made by looking at the structure correlations. These can be interpreted much like factor loadings in factor analysis.

MANOVA is used when you have an experiment with several related dependent measures. The analysis is essentially an extension of discriminant analysis to experimental data. Interpretation is based on the significance of the discriminant functions extracted for each effect, the Roy–Bargman stepdown test, and the structure correlations.

MANOVA also can be used to analyze data from a within-subjects design. Because of the restrictive assumptions of the univariate within-subjects ANOVA and the controversy surrounding error term selection for that analysis, MANOVA should be considered as an alternative. In MANOVA, the repeated measures taken from each subject are treated as correlated dependent variables. Using MANOVA in this capacity circumvents many of the problems associated with the traditional within-subjects ANOVA.

Multiway frequency analysis is a nonparametric multivariate statistic with a variety of applications. It can be used to analyze categorical data from an experiment or categorical variables from a correlational study. It also can be used on interval or ratio data in instances in which your data do not meet the assumptions of the analysis of variance. One form of multiway frequency analysis commonly used is loglinear analysis.

Path analysis is used to test a clearly specified causal model. Using a theory, you develop a causal model, measure your variables, and then use a series of simple multiple regression analyses to derive path coefficients. The path coefficients are used as estimates of the magnitude of causal relationships among variables. Interpretation is facilitated by looking at both direct and indirect effects of variables. Causal models can be of several types. Simple causal chains propose that there is a linear path from one variable to another. More complex models can involve complex path linkages. Models can be either recursive (which contain no causal loops) or nonrecursive (which contain causal loops). Conceptually, recursive models are easier to analyze and interpret. Structural equation modeling is a technique related to path analysis. It differs from path analysis in that it allows you to look at hypothetical variables called latent variables in your model. Finally, multivariate analyses are complex and tricky to use. Don't try to use them until you have a sound understanding of how they work, what assumptions they make, and how results can be interpreted.

KEY TERMS

univariate strategy
multivariate strategy
multivariate design
multiple regression
discriminant analysis
canonical correlation
factor analysis
multivariate analysis of variance (MANOVA)
partial correlation
part correlation
multiple *R*
R-square
beta weight
multiway frequency analysis
loglinear analysis
path analysis
structural equation modeling (SEM)
latent variable

16

CHAPTER

Reporting Your Research Results

CHAPTER OUTLINE

APA Writing Style

Writing an APA-Style Research Report
- Getting Ready to Write

Parts and Order of Manuscript Sections
- The Title Page
- The Abstract
- The Introduction
- The Method Section
- The Results Section
- The Discussion Section
- The Reference Section
- Footnotes
- Tables
- Figures

Elements of APA Style
- Citing References in Your Report
- Citing Quoted Material
- Using Numbers in the Text
- Avoiding Biased Language

Expression, Organization, and Style
- Precision and Clarity of Expression
- Economy of Expression
- Organization
- Style
- Making It Work
- Avoiding Plagiarism and Lazy Writing

Telling the World About Your Results
- Publishing Your Results
- Paper Presentations
- The Ethics of Reporting or Publishing Your Results

Summary

Key Terms

Your journey through the world of research has taken you through the steps involved in choosing a research question, developing hypotheses, choosing a general strategy and specific design to test your hypotheses, and describing and analyzing your data. The final step in this process is to tell the world what you did and what you found.

Reporting your research results is perhaps the most important step because it is only by this reporting that science progresses. Other scientists working in the field need to know what you have done: the questions that you have asked, the methods that you have used to address them, and the answers that you have found. This step is not only essential for progress but also required to assess the reliability of your findings and the soundness of your conclusions. Only when your research has been reported can others attempt to replicate and extend your findings.

Effectively communicating the results of your research requires you to know what must be said and how to communicate it clearly and in the proper format. This chapter discusses how to organize and present your research findings. Even if you are not planning to pursue a career in psychological research, you will find that much of the information contained in this chapter is valuable. Many occupations—especially those at a managerial or technical level—require you to organize facts, to draw conclusions, and to present the facts and conclusions clearly and logically in a written report. Although the format and contents of your reports are likely to differ from those of the scientific report described here, the general principles of organization and composition will be the same.

APA WRITING STYLE

Scientific journals in all disciplines specify the format, or *writing style*, that articles submitted to the journal are to follow. This writing style determines what subsections will be present in the report, how to present figures and tables, what rules are to be followed in typing the manuscript, and other such details.

In psychology, most journals follow the style established by the American Psychological Association (APA) in the *Publication Manual of the American Psychological Association* (6th ed., 2010). We refer to this manual simply as the APA publication manual. When you follow the manual, your manuscript conforms to APA style.

Sometimes you may come across a journal that does not completely follow APA style. Nevertheless, these journals usually follow APA style closely with only a few minor exceptions. (These exceptions are usually indicated on or near an inside cover of each journal issue under a heading such as "Notes to Authors.") Therefore, most of what you learn about APA style will apply even to non-APA psychology journals. The following discussion provides a condensed description of how to write a research report in APA style. We cannot present a comprehensive review of APA style in the space available here. (For additional details, consult the APA publication manual.) In this chapter, we also discuss how to avoid some common errors of composition and grammar and how to prepare a conference presentation.

WRITING AN APA-STYLE RESEARCH REPORT

There are seven main sections in an APA-style research report manuscript: the title page, abstract, introduction, method, results, discussion, and references. There are many guidelines to follow when you are preparing an APA-style manuscript. For example, specific instructions govern margins, line spacing, the content of each section, and how to present information in tables and figures. In this section, we discuss some of the basic guidelines for preparing an APA-style manuscript.

Getting Ready to Write

For this discussion, we assume that you will be creating your manuscript by using word processing software (such as Microsoft Word or Corel WordPerfect). Before you begin to type, make sure that the settings of your word processor conform to the requirements of an APA-style manuscript. In most cases, you can use the default settings, but there may be exceptions that will require a change. In particular, you should pay attention to the settings for paper size, font (typeface and size), margins, line spacing, hyphenation, and justification.

To create an APA-style manuscript, select the 8½ × 11-inch paper size (usually the default). As for typeface, APA prefers Times New Roman. A *serif* typeface such as this is preferred for the body of the manuscript (i.e., typeface that has lines extending from the beginnings and ends of letters) because it is easier to read. You may use a *sans serif* (without the lines) such as Arial for figures because this typeface gives a very clean, clear look. Set the font size to 12 points. Be sure to turn off any setting that compresses the typeface or reduces the spacing between characters.

Use double spacing throughout your manuscript. This goes for the spaces between lines of text, between lines in the title, and after headings, footnotes, quotations, references, figure captions, and all elements of a table (although you may use single spacing in tables or figures). You may use more space before or after an equation, but you should *not* use single or one-and-a-half spacing in your manuscript, except in some cases in figures and tables (APA, 2010). Spacing after punctuation marks varies according to where a punctuation mark occurs. Use a single space after commas, colons, and semicolons. Use a single space after periods appearing in references and after initials. Use double space after periods that end sentences.

Be sure to turn off right justification so that the last words of the lines on a page form a ragged right edge. Also, check to be sure that you disable automatic hyphenation. APA style requires that a word too long to fit on a given line appear in its entirety on the next line. This will happen only if automatic hyphenation is off.

All margins for an APA-style manuscript must be at least 1 inch all around (top, bottom, left, and right). With most word processors, 1-inch margins are usually the default. If they are not, consult your program's help files to find out how to change the margins. The length of a typed line has a maximum of 6½ inches.

Print the completed manuscript on 8½- × 11-inch heavy white bond paper. Use a high-quality printer (e.g., inkjet or laser printer) and check to be sure that the manuscript is free of defects such as light printing.

Formatting a Page Each page of an APA manuscript includes a **running head** and a page number The running head is a shortened version of your title having no more than 50 characters (including punctuation and spaces). Type the running head flush left at the top of each page of your manuscript. On the title page only, type "Running head:" by using capital and lowercase letters followed by your running head typed in all capital letters (see Figure 16-1). Type the page number on the same line as your running head and to its right (the APA manual does not specify a number of spaces to leave between your running head and page number). The running head and page number serve as the manuscript page header for each page. On all subsequent pages, *do not* type the words "Running head." Rather, type only the running head itself at the top of each page (see Figure 16-2). Use your word processor's page header function to create the text and set up automatic page numbering so that the running head and page number appear on every page of your manuscript automatically.

Indent the first line of all paragraphs and footnotes (with the exception of the abstract, the first paragraph of block quotes, titles and headings, table titles and notes, and figure captions). Set the tabs on your word processor to give an indent of from five to seven spaces (or one-half inch).

Heading Structure Headings within a manuscript identify different sections and subsections. In an APA-style manuscript, you can have anywhere from one to five levels of headings. The structure for these five levels is as follows (APA, 2010, p. 62):

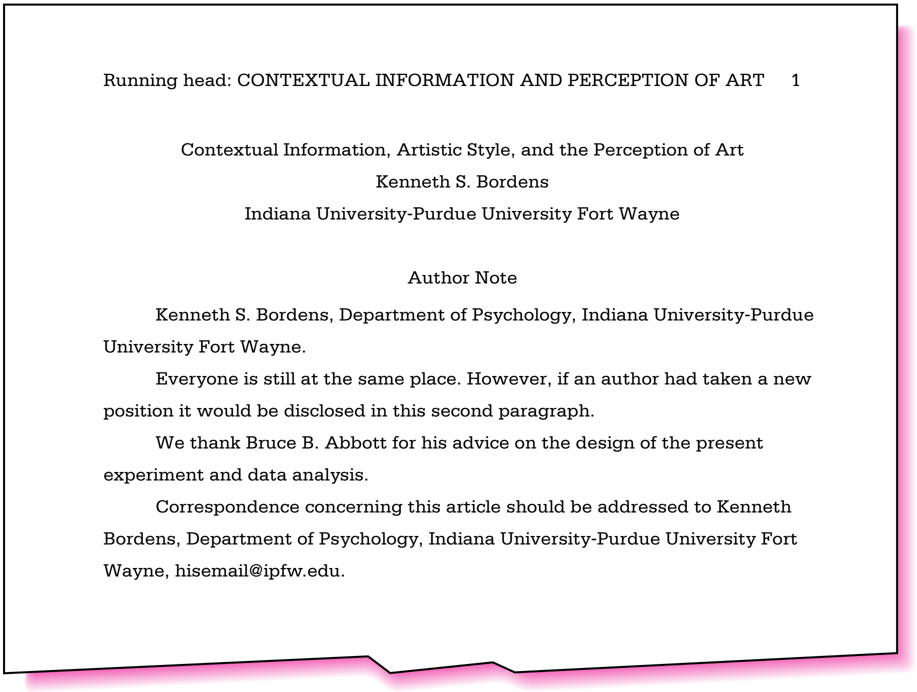
Centered, boldface, upper and lowercase (Level 1)

Flush left, boldface, upper and lowercase (Level 2)

Indented, boldface lowercase paragraph heading ending with a period. (Level 3)

Indented, boldface italicized, lowercase paragraph heading ending with a period. (Level 4)

Indented, italicized, lowercase paragraph heading ending with a period. (Level 5)



Running head: CONTEXTUAL INFORMATION AND PERCEPTION OF ART 1

Contextual Information, Artistic Style, and the Perception of Art
Kenneth S. Bordens
Indiana University-Purdue University Fort Wayne

Author Note

Kenneth S. Bordens, Department of Psychology, Indiana University-Purdue University Fort Wayne.

Everyone is still at the same place. However, if an author had taken a new position it would be disclosed in this second paragraph.

We thank Bruce B. Abbott for his advice on the design of the present experiment and data analysis.

Correspondence concerning this article should be addressed to Kenneth Bordens, Department of Psychology, Indiana University-Purdue University Fort Wayne, hisemail@ipfw.edu.

FIGURE 16-1 Sample title page.

In most cases, you will use only a three-level heading structure. However, you may use the other levels depending on the needs of your paper. A three-level heading structure looks like this:

Method (Level 1)

Participants (Level 2)

Participants not meeting requirements. (Level 3)

For manuscripts requiring more than three levels of headings, consult the APA publication manual (2010, pp. 62–63).

QUESTIONS TO PONDER

1. How do you set up a paper using APA writing style?
2. What is the heading structure used in an APA-style manuscript?

PARTS AND ORDER OF MANUSCRIPT SECTIONS

Each section of an APA-style manuscript includes important information that informs your reader what your paper is about, how you conducted your study, what you found, how your results relate to previous research and theory, and a list of references. In the sections that follow, we present information about each major section of an APA-style. The order in which we present these sections is the order in which they should appear in your paper.

The Title Page

The **title page** includes (in order) your running head and page number (on the title page remember to type the words "Running head" as part of the page's running head), the title of your paper, author name, the author's institutional affiliation, and any author notes. Place the title, author, and institutional affiliation information on the top half of the title page centered between the left and right margins. Figure 16-1 shows how to format a title page for a single author. Consult the APA publication manual (2010) to see how to format a title page for other types of authorship.

Title When researchers looking for relevant articles on a particular topic scan the table of contents of a journal or an abstract in PsycINFO, the title of an article first captures attention. If the title fails to communicate clearly what the paper is about, readers may skip the paper.

An unread paper is useless. To avoid this fate for your paper, make your title concise yet informative. Avoid using words that add little to the meaningfulness of your title (e.g., "An Experimental Investigation of . . ." and "A Correlational Field Study of . . ."). Keep your title short enough to avoid confusion about your research, but not so short that it fails to convey the topic of your paper. The recommended length for a title is no more than 12 words.

Taking our cue from the story of "Goldilocks and the Three Bears," we present the following examples showing a title that is too long, a title that is too short, and one that is just right:

> *Too long:* An Experimental Study of the Effect of Delay of
> Reinforcement on Discrimination Learning in White Rats
>
> *Too short:* The Effect of Reinforcement on Learning
>
> *Just right:* Effect of Delay of Reinforcement on Discrimination
> Learning in Rats

In the first example, the words "An Experimental Study of " and "White" add extraneous words to your title. The second title is too general. A potential reader has only a vague idea about the focus of your study. The third example concisely conveys the essence of your study.

Type your title in the specified position on the title page. If multiple lines are required, double-space them just as you do other text. Capitalize the first letter of the first word and of all subsequent words (except for articles, prepositions of three letters or less, and conjunctions).

Author Name(s) and Affiliation(s) If you are the sole author of the paper, your name goes one double-spaced line beneath the title. Include your given name, middle initial(s), and last name (in that order), centered between the margins. To avoid confusion you should use the same format for your name for all papers you prepare. Do not include any titles (such as Mr., Ms., Dr.) or degrees (B.A., M.A., Ph.D., M.D., etc.).

Your affiliation identifies where you were when you conducted your research. This is usually the organization that provided the local facilities and/or support for your research (usually a university or college). Its name appears one double-spaced line below yours on the title page, centered between the margins.

If there are two or more authors, how you organize the information depends on whether everyone has the same affiliation. Consult the APA publication manual for details. With multiple authors, the APA publication manual directs that multiple authors must be listed in order of their degree of contribution to the paper. If all authors contributed equally, they should work out some method for listing (e.g., alphabetically). If you do this, you should note it in the manuscript in an "author note."

Author Note Published APA-style journal articles include an **author note**, which is a small footnote at the bottom of the first page to identify each author's departmental affiliation, provide acknowledgments, state disclaimers or conflicts of interest, and indicate how readers can contact the author. Type the author note on the title page. Author notes are not numbered, nor should you refer to them in the body of your paper.

Arrange the author note itself in four paragraphs (APA, 2010). The first paragraph identifies the author's affiliation when the study was conducted. The second paragraph lists any changes in author affiliation since the study was conducted. The third paragraph includes acknowledgments such as grants that supported the research or the names of those who made special contributions to the research (such as informal review, assistance with the research design, or statistical analysis). It also indicates any special circumstances concerning the paper (such as that the study was a replication of an earlier one or that the study was done as a requirement for a degree). The fourth paragraph presents the "point of contact." Here you provide a complete mailing address (including, if desired, an e-mail address at the end of the paragraph) where readers can contact you. An author note is also used to indicate any conflicts of interest or biases an author may have.

The Abstract

The **abstract** is a concise summary of your paper. Each journal has its own requirements concerning the length of your abstract. In most cases, the length of your abstract will be between 150 and 250 words (APA, 2010, p. 27). The content of your abstract will depend on the nature of your paper. In the abstract for an empirical study, include the following information (APA, 2010, p. 26):

1. Information on the problem under study (preferably in one sentence)
2. The nature of the subject sample (e.g., age and sex)
3. A description of the methods used, including equipment, procedures for gathering data, names of tests, and so on
4. A statement of the findings, including information on levels of statistical significance, effect sizes, and confidence intervals
5. A statement of the conclusions drawn and any implications or applications of your results

Although short, the abstract is important. Journals include abstracts of papers at the beginning of the final journal article and in PsycINFO entries. A potential reader will use your abstract to decide whether to read your paper as part of a literature search. If your abstract is poorly written, readers may fail to understand the significance of your work and may pass it by. Thus, you should put effort into writing a clear, concise abstract.

Your abstract is the first substantive section of your paper. However, you typically write it *after* you have written the rest of your paper, when you will have a clearer idea about what must be included in the abstract.

The APA manual defines four qualities that make for a good abstract. First, your abstract must be *accurate*. This means that the information in your abstract reflects what is in the body of your paper. Do not include any information in the abstract that does not appear in your paper. Second, your abstract should be *nonevaluative*. You should report on your study and avoid adding any comments on what is in your paper. Third, your abstract should be *coherent and readable*. Write your abstract using clear and concise language. Generally, write in the active rather than passive voice and write in the present tense (except when describing specific manipulations or results). Fourth, make your abstract as *concise* as possible. Include only the most important information and write in concise sentences. Remember, you have a limited number of words, so make them count.

Formatting the Abstract Type your abstract on a separate page, immediately after your title page and number it as page 2. Type the word "Abstract" centered on the line following your running head and page number. On the next line, begin your abstract. Do *not* indent the first line of the abstract. Include a list of keywords, centered immediately below your abstract (see Figure 16-2).

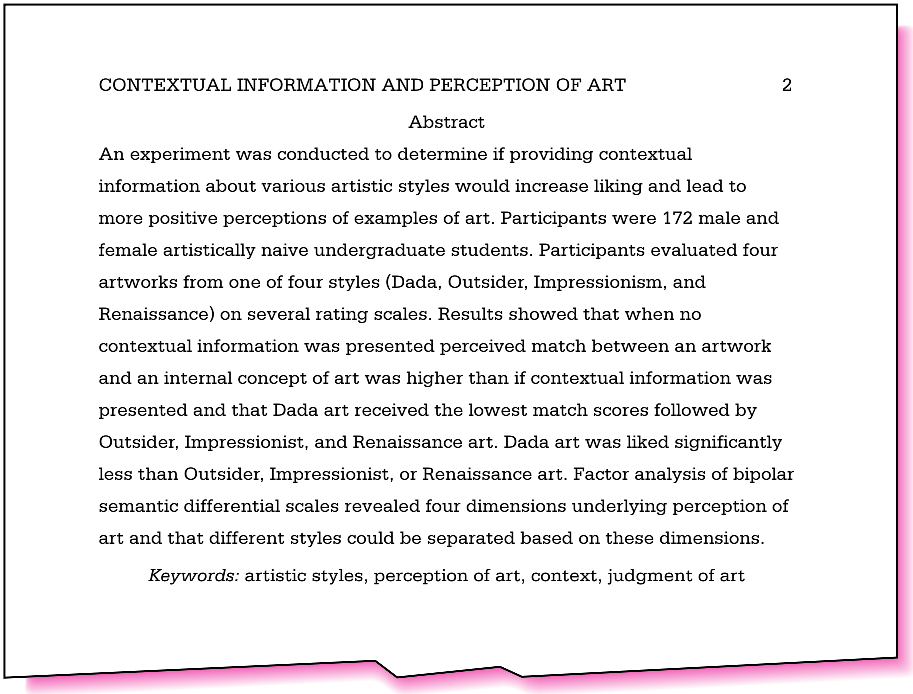
QUESTIONS TO PONDER

1. What information is included on the title page, and in what order would you find that information (from the top of the page to the bottom of the page)?
2. What is an abstract, and why is it so important?
3. What information goes into an abstract, and how long should an abstract be?

The Introduction

The text of the paper begins with the **introduction**. The primary function of the introduction is to describe the problem studied and your basic research strategy. Before writing your introduction, the APA manual suggests asking yourself the following questions (APA, 2010, p. 27):

1. Why is the issue studied important?
2. How does your study relate to previous research in the area and how does it differ from other studies on the same issue?
3. What are the hypotheses and objectives of your study and how do they relate to relevant theory (if they do)?



CONTEXTUAL INFORMATION AND PERCEPTION OF ART 2

Abstract

An experiment was conducted to determine if providing contextual information about various artistic styles would increase liking and lead to more positive perceptions of examples of art. Participants were 172 male and female artistically naive undergraduate students. Participants evaluated four artworks from one of four styles (Dada, Outsider, Impressionism, and Renaissance) on several rating scales. Results showed that when no contextual information was presented perceived match between an artwork and an internal concept of art was higher than if contextual information was presented and that Dada art received the lowest match scores followed by Outsider, Impressionist, and Renaissance art. Dada art was liked significantly less than Outsider, Impressionist, or Renaissance art. Factor analysis of bipolar semantic differential scales revealed four dimensions underlying perception of art and that different styles could be separated based on these dimensions.

Keywords: artistic styles, perception of art, context, judgment of art

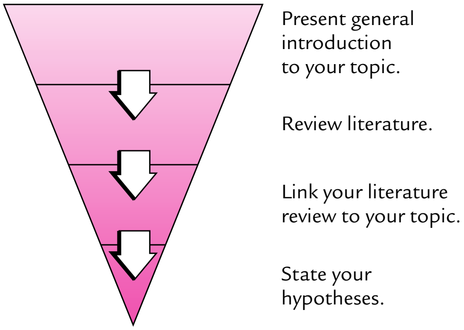
FIGURE 16-2 Sample abstract.

4. How do your hypotheses relate to your research design?
5. What are the theoretical and practical implications of your study?

Your introduction should include three essential elements (APA, 2010): an exploration of the importance of the problem examined, a description of relevant previous research and theory, and a clear statement of your hypotheses and how they relate to your research design. To help the reader understand why you conducted your study, your introduction should include the following information:

1. An introduction to the topic under study
2. A brief review of the research findings and theories related to the topic
3. A statement of the problem to be addressed by the research (identifying an area in which knowledge is incomplete)
4. A statement of the purpose of the research (always to solve the problem identified but perhaps only a specific aspect of it)
5. A brief description of the research strategy, intended to establish the relationship between the question being addressed and the method used to address it
6. A description of any predictions about the outcome and of the hypotheses used to generate those predictions

FIGURE 16-3 General-to-specific organization of an APA-style introduction.


Present general introduction to your topic.
Review literature.
Link your literature review to your topic.
State your hypotheses.

To provide this information in a comprehensible way, the structure of the introduction proceeds from the general to the specific. The inverted triangle shown in Figure 16-3 illustrates this structure. In the opening paragraph of your introduction, discuss (in general terms) the issue that you have chosen to study. Next, develop the underlying logic and rationale for your study in more specific terms by reviewing relevant research and integrating its findings. Then identify the problem addressed by your research and state the purpose of your study. Finally, show how your study addresses the question and state your specific hypotheses.

If your introduction includes information of a controversial nature, you should present this information in a fair and balanced manner. You should avoid expressing strong opinions on one side or another of the controversial issue. If you must express a personal opinion in the introduction, you should offer it without hostility and without making personal attacks on those with whom you disagree. When using citations to support your view or your research, you must present the research fairly and not out of context (APA, 2010).

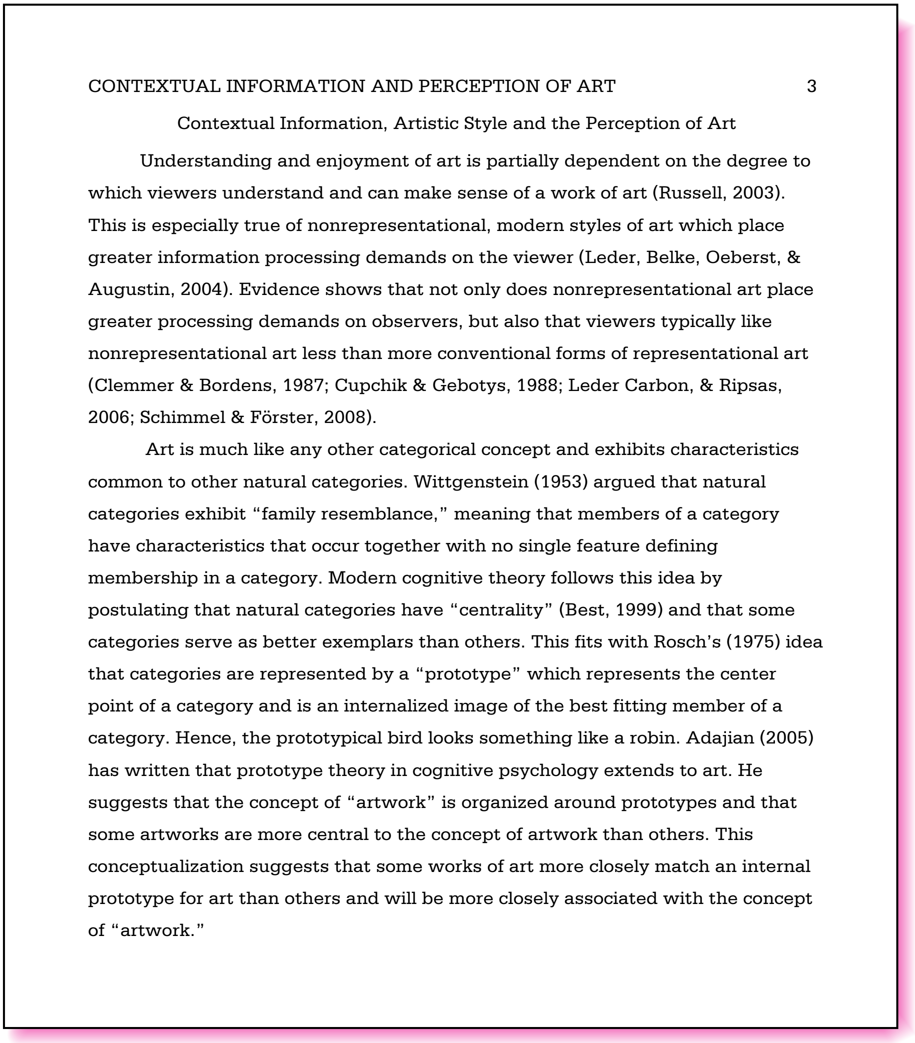
Students often have difficulty determining what should go into the literature review of previous findings. Should it include a comprehensive review of *all* relevant research or be limited to a few papers that relate specifically to your research? The answer: Try for something between these extremes.

You can assume that your reader has some knowledge of the basic psychological concepts that underlie your study. Your job is to bring your reader up to date on the literature that most directly relates to your study. For example, if you investigated the effect of amount of reinforcement on behavior modification of developmentally delayed children, you need not review all research on operant conditioning and reinforcement. You can assume that your reader has some knowledge of the basic concepts of operant conditioning. Instead, focus on the important issues relating directly to using operant conditioning to modify the behavior of developmentally delayed persons.

Assume that you have decided to focus your introduction on the important issues. You then head to the library and, to your shock and horror, you find 200 articles that relate in some way to your research topic. Just how many of these must be included in your literature review? Of course, you cannot hope to review them all.

In fact, such a comprehensive literature review is inappropriate for a research paper. Your review should focus on those issues that are most important for establishing the rationale of your study. Therefore, you should identify all the papers most directly relevant to the issues raised by your introduction. Within this narrower area, you can cite all or at least many of the relevant papers.

Formatting the Introduction To follow APA style, begin the introduction on a new page with your running head and page number (page 3) at the top of the page. Next, center the title of the paper at the top of the page and start the introduction immediately below the title. The title is the same one you used on the title page. Do *not* type the heading “Introduction.” In addition, neither your name nor your affiliation appears on the first page of the introduction. Figure 16-4 shows a sample



CONTEXTUAL INFORMATION AND PERCEPTION OF ART 3

Contextual Information, Artistic Style and the Perception of Art

Understanding and enjoyment of art is partially dependent on the degree to which viewers understand and can make sense of a work of art (Russell, 2003). This is especially true of nonrepresentational, modern styles of art which place greater information processing demands on the viewer (Leder, Belke, Oeberst, & Augustin, 2004). Evidence shows that not only does nonrepresentational art place greater processing demands on observers, but also that viewers typically like nonrepresentational art less than more conventional forms of representational art (Clemmer & Bordens, 1987; Cupchik & Gebotys, 1988; Leder Carbon, & Ripsas, 2006; Schimmel & Förster, 2008).

Art is much like any other categorical concept and exhibits characteristics common to other natural categories. Wittgenstein (1953) argued that natural categories exhibit “family resemblance,” meaning that members of a category have characteristics that occur together with no single feature defining membership in a category. Modern cognitive theory follows this idea by postulating that natural categories have “centrality” (Best, 1999) and that some categories serve as better exemplars than others. This fits with Rosch’s (1975) idea that categories are represented by a “prototype” which represents the center point of a category and is an internalized image of the best fitting member of a category. Hence, the prototypical bird looks something like a robin. Adajian (2005) has written that prototype theory in cognitive psychology extends to art. He suggests that the concept of “artwork” is organized around prototypes and that some artworks are more central to the concept of artwork than others. This conceptualization suggests that some works of art more closely match an internal prototype for art than others and will be more closely associated with the concept of “artwork.”

FIGURE 16-4 Sample introduction.

introduction. Notice how the author followed the general-to-specific structure suggested in Figure 16-4 by beginning with a statement of the problem to study, followed by a review of relevant literature. It ends with a statement of the hypothesis to be tested (not shown).

The Method Section

After you have established the rationale for your study and stated your hypotheses in the introduction, you then must tell your reader exactly how you conducted your study. You do this in the **method section**, which describes in detail the characteristics of your subjects, materials, and apparatus used, research design, as well as the procedures followed. The level of detail should be sufficient to allow another researcher to replicate your study. If your paper uses a methodology described before, you may give a brief summary of the methods used and refer the reader to the more detailed published account.

The method section is divided into subsections to improve organization and readability. APA style permits considerable flexibility in how you divide and label the various subsections. The most common format contains the following subsections: *participants* (or *subjects* if you used animals), *apparatus* (or *materials* if this descriptor is more appropriate), and *procedure*. If you consider it necessary, you also may include a *design* subsection to clarify your design for your readers. A *design section* is particularly useful when your study used an unconventional or complex design. You also can combine subsections if this improves the clarity of the report. A description of each subsection follows.

Participants or Subjects If humans participated as subjects in your study, describe them in a **participants subsection**. In this section, you specify the nature and size of the sample used in your study. Specify the number of participants and provide information on relevant demographic variables (such as sex, age, race, ethnicity), the procedures used for selection of participants and their assignment to treatments, any special agreements made with participants (such as payment for their participation), and information on personal characteristics of the participants that are relevant to your research (such as IQ and personality). Also, report any special characteristics of your participants, such as mental impairment, psychopathology, or special abilities.

If your subjects were animals, describe them in a **subjects subsection**. Describe the genus, species, strain, and any other important relevant information (such as the supplier). Also, give the number of animals used in the study and their sex, age, weight, and physical condition. Provide details of the care and housing of the animals (e.g., whether they were housed individually or in groups, whether they were given free access to food and water, and the scheduling of light and darkness in the colony room). In this section, you also specify the number of subjects assigned to each condition in your experiment and any other information (such as special handling) that your reader needs to know in order to replicate your study. Finally, in either subsection, indicate that you treated your participants or subjects in accordance with APA ethical codes.

Apparatus or Materials[1] You describe the equipment or any materials used to measure behavior in this section. If you used primarily equipment (e.g., slide projectors, operant chambers, computers), describe that equipment in an **apparatus subsection**. If you used primarily written materials (e.g., a questionnaire, summaries of criminal cases, or rating scales), describe them in a **materials subsection**. In either case, the level of detail necessary in your description depends on the nature of the equipment or materials used.

If you used a commercial piece of laboratory equipment (e.g., an operant chamber or a computer), you do not have to detail its characteristics. Instead, simply provide the name of the manufacturer and the model number of the equipment. Similarly, if you used a standardized test (such as the Stanford–Binet or the Bem Sex Roles Inventory), simply name the test (and the version, if relevant), and describe how it was obtained.

If you designed special equipment or developed a new measure, you must describe the equipment or measure in detail. If you designed a special operant chamber, for example, give its dimensions, materials of construction, and the types, characteristics, and locations of attached equipment (such as feeders, houselights, response levers, and sound sources). In short, provide any information that your reader would need to reproduce your chamber in its essential details. Similarly, describe any measures you developed (e.g., questionnaires).

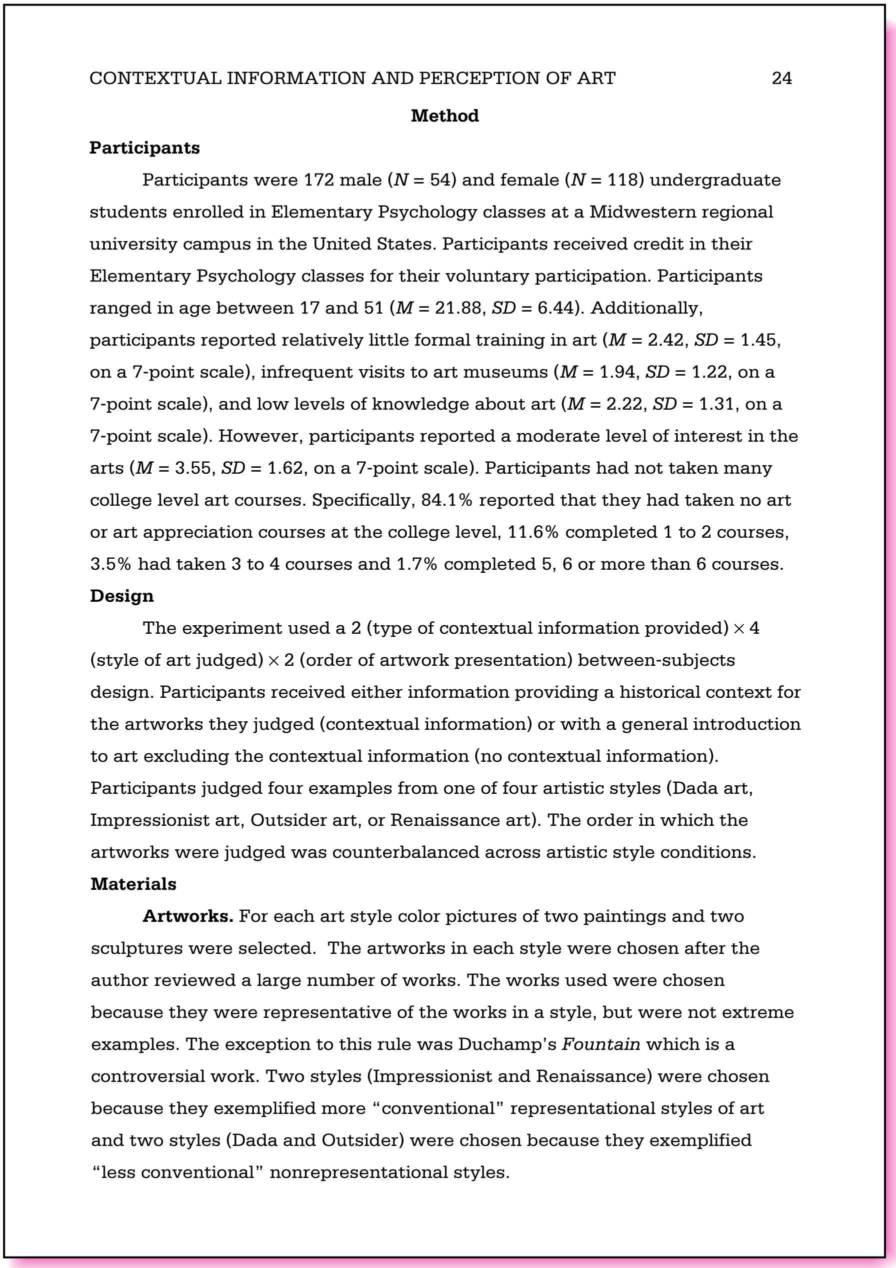
Although you should provide enough information to enable another researcher to replicate your study, it is not feasible to reproduce extensive materials (such as a 250-item questionnaire or lengthy instructions) in your method section. If you used such materials, inform your readers where and how they can obtain them. Some journals allow you to print such materials in an appendix.

Procedure In the **procedure subsection**, tell your reader precisely the procedure you followed throughout the course of the study. Describe the conditions to which subjects were exposed or under which they were observed, what behaviors were recorded, how the behaviors were measured or scored, when the measures were taken, and any debriefing procedures. Provide enough information about the procedure so that another researcher could reproduce its essential details.

If you used animal subjects, describe the following: how you handled the animals, the length of the experimental sessions, any special deprivation schedules, and to what manipulations the subjects were exposed. If humans were used, include details about the instructions they received (if you cannot reproduce them, describe them in detail), informed-consent procedures, procedures for assigning subjects to conditions, and how the experimental manipulations were introduced.

Formatting the Method Section The method section begins immediately after the end of the introduction (do not necessarily start it on a new page). Center the word "Method" (not "Methods," a common error we see often) as a Level 1 heading. On the next double-spaced line, type the word "Participants" beginning at the left margin as a Level 2 heading. Again, move down a double-spaced line, indent, and start the

[1] Although the APA manual (2010) no longer makes a formal distinction between apparatus and materials sections, it is still a useful distinction and we recommend using it.



CONTEXTUAL INFORMATION AND PERCEPTION OF ART 24

Method

Participants

Participants were 172 male (*N* = 54) and female (*N* = 118) undergraduate students enrolled in Elementary Psychology classes at a Midwestern regional university campus in the United States. Participants received credit in their Elementary Psychology classes for their voluntary participation. Participants ranged in age between 17 and 51 (*M* = 21.88, *SD* = 6.44). Additionally, participants reported relatively little formal training in art (*M* = 2.42, *SD* = 1.45, on a 7-point scale), infrequent visits to art museums (*M* = 1.94, *SD* = 1.22, on a 7-point scale), and low levels of knowledge about art (*M* = 2.22, *SD* = 1.31, on a 7-point scale). However, participants reported a moderate level of interest in the arts (*M* = 3.55, *SD* = 1.62, on a 7-point scale). Participants had not taken many college level art courses. Specifically, 84.1% reported that they had taken no art or art appreciation courses at the college level, 11.6% completed 1 to 2 courses, 3.5% had taken 3 to 4 courses and 1.7% completed 5, 6 or more than 6 courses.

Design

The experiment used a 2 (type of contextual information provided) × 4 (style of art judged) × 2 (order of artwork presentation) between-subjects design. Participants received either information providing a historical context for the artworks they judged (contextual information) or with a general introduction to art excluding the contextual information (no contextual information). Participants judged four examples from one of four artistic styles (Dada art, Impressionist art, Outsider art, or Renaissance art). The order in which the artworks were judged was counterbalanced across artistic style conditions.

Materials

Artworks. For each art style color pictures of two paintings and two sculptures were selected. The artworks in each style were chosen after the author reviewed a large number of works. The works used were chosen because they were representative of the works in a style, but were not extreme examples. The exception to this rule was Duchamp's *Fountain* which is a controversial work. Two styles (Impressionist and Renaissance) were chosen because they exemplified more "conventional" representational styles of art and two styles (Dada and Outsider) were chosen because they exemplified "less conventional" nonrepresentational styles.

FIGURE 16-5 First page of a method section. The method section begins immediately after the end of the introduction, not necessarily on a new page.

first paragraph of the participants (or subjects) subsection. Follow the same format you used with the participants subsection for the apparatus (or materials) and procedure subsections. Figure 16-5 shows an example method section with participants, materials, and procedure subsections. Note how each subsection contributes to your understanding of the described experiment.

QUESTIONS TO PONDER

1. What information is included in the introduction to an APA-style paper? How is the introduction organized?
2. What information would you expect to find in the method section?
3. Describe the various subsections of the method section.

The Results Section

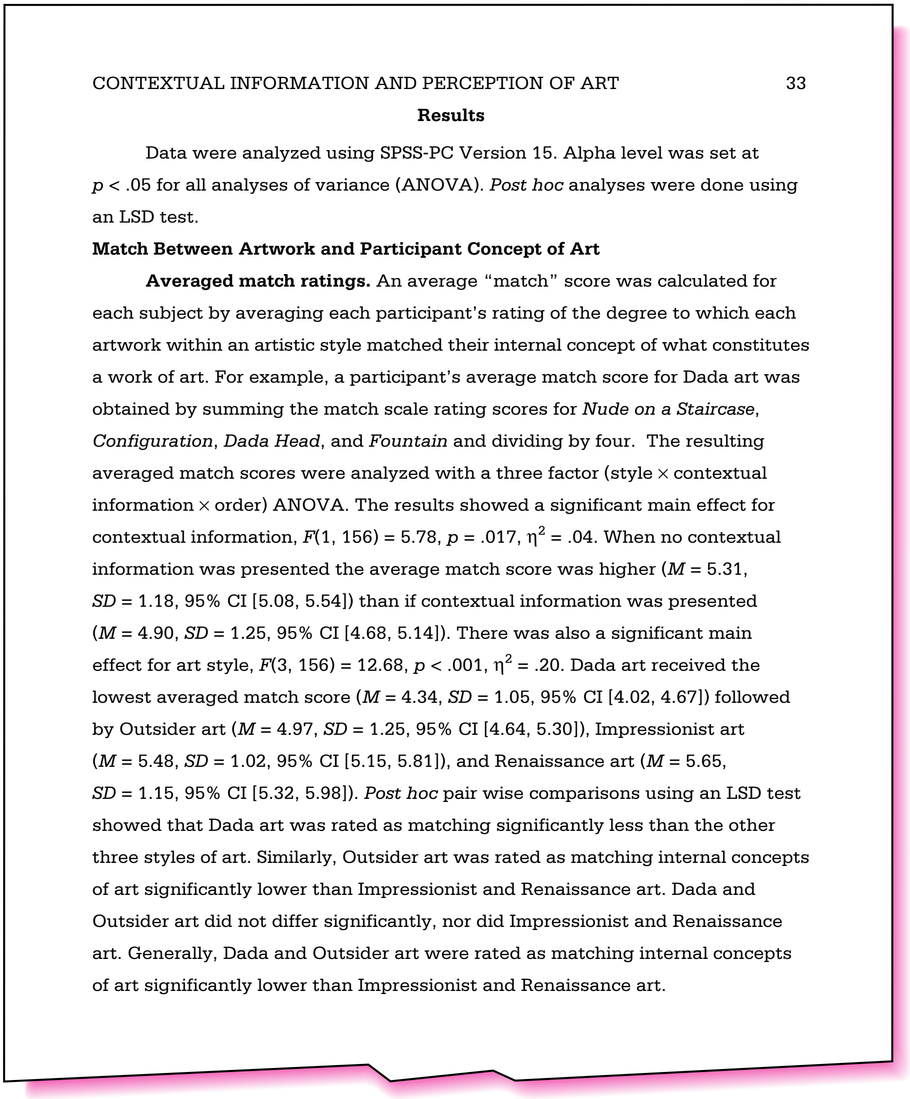
The purpose of your **results section** is to report your findings. You should present all relevant data summaries and analyses. As a rule, do not present raw (unanalyzed) or individual data unless the focus of the study was on the behavior of individual subjects (e.g., case history or single-subject design). If your analysis is complex, you may want to provide an overview of your strategy for data analysis in the opening paragraph. Outline for the reader which statistical tests you applied and in what order.

Your results section should be primarily a narrative where you describe what you found. Make this narrative the driving force behind your results section. The results of descriptive and inferential statistics also will appear in your results section. However, these statistics should support the narrative statements that you make. Too often, students allow statistics to drive the discourse in the results section, throwing in everything but the kitchen sink. What results is a compilation of numbers with little coherence. Remember, you should verbally describe what you found and include references to relevant statistics to support what you say.

If you are using a statistical test that is not generally available, indicate to your reader where you found the test and how your reader can obtain information about it. Next, report the results of any tests used to establish that your data met the requirements of the applied statistical tests (e.g., homogeneity of variance and normality). Report any data transformations that you applied to your data.

After you have presented this preliminary information, you can report your results. Include values of any descriptive (e.g., means and standard deviations) and inferential statistics (e.g., t-tests, analyses of variance) that you calculated, along with the relevant *p* values. The APA manual strongly suggests reporting confidence intervals (95% or 99%) and effect size statistics for all significant effects. Do not interpret or discuss your findings in the results section (APA, 2010); you do this in the next section of your paper.

Formatting the Results Section The results section begins immediately after the method section on the same page as the end of the method section, if there is room. The results section should be a continuation of your paper. Center the heading "Results" (Level 1 heading), double-space, and indent to start the first paragraph of your new section. Figure 16-6 shows the first page of an example results section.



CONTEXTUAL INFORMATION AND PERCEPTION OF ART 33

Results

Data were analyzed using SPSS-PC Version 15. Alpha level was set at $p < .05$ for all analyses of variance (ANOVA). *Post hoc* analyses were done using an LSD test.

Match Between Artwork and Participant Concept of Art

Averaged match ratings. An average "match" score was calculated for each subject by averaging each participant's rating of the degree to which each artwork within an artistic style matched their internal concept of what constitutes a work of art. For example, a participant's average match score for Dada art was obtained by summing the match scale rating scores for *Nude on a Staircase*, *Configuration*, *Dada Head*, and *Fountain* and dividing by four. The resulting averaged match scores were analyzed with a three factor (style × contextual information × order) ANOVA. The results showed a significant main effect for contextual information, $F(1, 156) = 5.78$, $p = .017$, $\eta^2 = .04$. When no contextual information was presented the average match score was higher ($M = 5.31$, $SD = 1.18$, 95% CI [5.08, 5.54]) than if contextual information was presented ($M = 4.90$, $SD = 1.25$, 95% CI [4.68, 5.14]). There was also a significant main effect for art style, $F(3, 156) = 12.68$, $p < .001$, $\eta^2 = .20$. Dada art received the lowest averaged match score ($M = 4.34$, $SD = 1.05$, 95% CI [4.02, 4.67]) followed by Outsider art ($M = 4.97$, $SD = 1.25$, 95% CI [4.64, 5.30]), Impressionist art ($M = 5.48$, $SD = 1.02$, 95% CI [5.15, 5.81]), and Renaissance art ($M = 5.65$, $SD = 1.15$, 95% CI [5.32, 5.98]). *Post hoc* pair wise comparisons using an LSD test showed that Dada art was rated as matching significantly less than the other three styles of art. Similarly, Outsider art was rated as matching internal concepts of art significantly lower than Impressionist and Renaissance art. Dada and Outsider art did not differ significantly, nor did Impressionist and Renaissance art. Generally, Dada and Outsider art were rated as matching internal concepts of art significantly lower than Impressionist and Renaissance art.

FIGURE 16-6 First page of a results section. The results section begins immediately after the end of the method section, not necessarily on a new page.

The results section is where you discuss any tables or figures that present data from your study. Although these will appear in the body of the *published* text, they do *not* appear in the body of the manuscript. Instead, you refer to the figure or table at the appropriate place. However, when referring to a figure or table, do not refer to its position. For example, *do not* say, "Figure 1, shown above, illustrates . . ." because the figure may not appear where you expected it to be placed in the published article. Simply refer to the figure by number in your manuscript: "Figure 1 shows the relationship between . . ."

Presenting the results of a statistical test is a bit tricky but not difficult once you become familiar with the process. You usually report the results of a statistical test in sentence format. The sentence states the effect being evaluated, whether or not the difference between treatment levels was statistically significant, the critical statistic used, the degrees of freedom, the value obtained for the statistic, the level of significance achieved, the measure of effect size used, and the effect size. Next, you provide examples of how to report the results from an analysis of variance in the body of your paper. Table 16-1 presents examples of how to report other statistics. In cases in which your analysis is very complex, you could present the results of your statistical tests in a table.

When reporting the results from an inferential statistical analysis, it is a good idea to begin by stating the alpha used to evaluate statistical significance. For example, you might say:

> All statistical tests employed an alpha level of .05.

Alternatively, you can indicate the alpha level when you report the results of your statistical analysis;

> At the .05 alpha level, the main effect of stimulus complexity was statistically significant, $F(2, 35) = 12.45$, $p = .034$, $\eta^2 = .06$.

When reporting the actual values of your statistical tests, include the value obtained for the test (e.g., the *F*-value), degrees of freedom, *p*-value, and an effect size statistic (e.g., eta squared, abbreviated η^2). When reporting *p*-values, report the exact value obtained from your computer printout to two or three decimal places. You should report *p*-values that are lower than .001 as $p < .001$ (and not for example, $p < .00001$). The following example shows how you should report the results of an inferential statistical test if you have previously stated your chosen alpha level:

> The main effect of stimulus complexity was statistically significant, $F(2, 35) = 12.45$, $p = .032$, $\eta^2 = .06$.

TABLE 16-1 Commonly Used Statistical Citations

STATISTIC	FORMAT
Analysis of Variance	$F(1, 85) = 5.96$, $p = .026$, $\eta^2 = .06$.
Chi-square	$\chi^2(3, N = 100) = 11.34$, $p < .001$
t Test	$t(56) = 4.78$, $p = .013$
z Test	$z = 2.04$, $p = .033$
Pearson correlation	coefficient $r = .87$ or $r = -.87$
Mean	M = 6.56
Standard deviation	*SD* = 1.96
Confidence interval	95% CI [5.67, 7.98]

Note: Numbers in parentheses are the degrees of freedom. For the analysis of variance, the first number in the parentheses is the degrees of freedom for the numerator and the second number the degrees of freedom for the denominator (error).

In your results section you must report descriptive statistics associated with any significant effects you found. You should report the means, standard deviations, and confidence intervals associated with each statistically significant effect. You have two options for reporting these statistics. You can organize descriptive statistics in a table. Use this option if you have many means relating to a complex interaction. The other option is to report the descriptive statistics in the body of your report. In this case, use the following format:

> Participants in the simple stimulus group identified more stimuli correctly (M = 12.56, *SD* = 2.43, 95% CI [11.2, 14.51]) than participants in the complex stimulus group (M = 9.24, *SD* = 2.66, 95% CI [7.32, 11.22]).

Putting the previous two example sentences together, your results section might contain the following:

> The main effect of stimulus complexity was statistically significant, $F(2, 35) = 12.45$, $p = .032$, $\eta^2 = .06$. Participants in the simple stimulus group identified more stimuli correctly (M = 12.56, *SD* = 2.43, 95% CI [11.2, 14.51]) than participants in the complex stimulus group (M = 9.24, *SD* = 2.66, 95% CI [7.32, 11.22]).

As you can see from these examples, APA style follows a consistent format for reporting statistical results. For statistical symbols, use the normal typeface for Greek letters and for acronyms (e.g., ANOVA) but italicize all other symbols that use standard alphabetical characters (e.g., *F*, *df*, *p*). Table 16-2 shows the abbreviations used for common statistics.

TABLE 16-2 Abbreviations for Statistical Symbols

ABBREVIATION	MEANING
CI	Confidence Interval
df	Degrees of freedom
F	*F* ratio
M	Arithmetic average (mean)
N	Number of subjects in entire sample
n	Number of subjects in limited portion of a sample
p	*p* value
SD	Standard deviation
t	*t* statistic
z	Results from *z* test or a *z* score
μ	Population mean (mu)
α	Alpha level
β	Beta
d	Cohen's *d* (effect size statistic)
η^2	eta squared (effect size statistic)
ω^2	omega squared (effect size statistic)

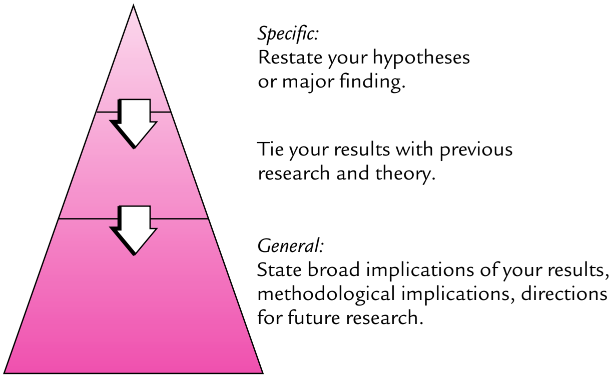

Specific:
Restate your hypotheses or major finding.
Tie your results with previous research and theory.
General:
State broad implications of your results, methodological implications, directions for future research.

FIGURE 16-7 Specific-to-general organization of an APA-style discussion section.

The Discussion Section

In the **discussion section**, you interpret your results, draw conclusions, and relate your findings to previous research or theory. The structure of your discussion section, as shown in Figure 16-7, reverses that of the introduction; rather than moving from general to specific, it moves from specific research findings to general implications.

Begin your discussion section with a brief restatement of your hypotheses. Next, briefly indicate whether your data were consistent with your pre-experimental hypotheses. Use the remainder of the discussion section to integrate your findings with previous research and theory. Discuss how consistent your findings are with previous work in the area.

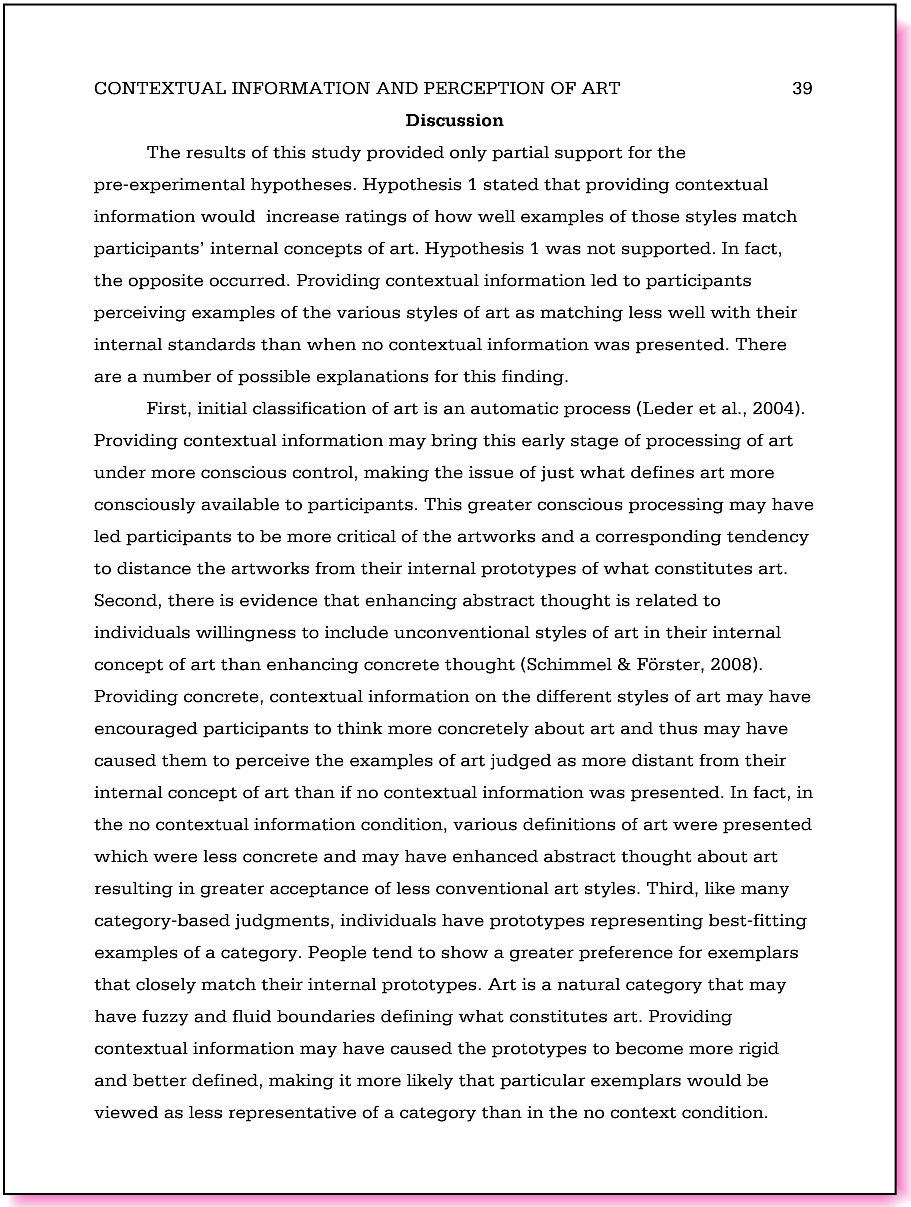
If your study yielded results that are discrepant from previous work, you should speculate on why the discrepancies emerged. Also, point out any problems encountered during the course of your research that might temper any conclusions drawn from your study. You should report on any methodological problems that became evident when you actually ran your study. Finally, indicate what implications your research has for future research in the area. Point out any specific areas that need to be investigated further.

In the discussion section, you have license to speculate on the importance of your findings. Avoid the temptation to overstep the bounds of that license. You must base your interpretations on your data, data from previous research, and/or established theory. Avoid the temptation to make unsubstantiated interpretations even if they make intuitive sense to you.

Figure 16-8 shows an example of the first page of a discussion section. The author of this paper followed the specific-to-general organization shown in Figure 16-7. Notice that the author begin with a brief statement of what was found and follow by integrating their findings with those of other research in the area.

QUESTIONS TO PONDER

1. What would you expect to find in the results section of a manuscript?
2. How is the results section formatted, and how are statistics reported?
3. How is the discussion organized, and what would you expect to find in the discussion section?



CONTEXTUAL INFORMATION AND PERCEPTION OF ART 39

Discussion

The results of this study provided only partial support for the pre-experimental hypotheses. Hypothesis 1 stated that providing contextual information would increase ratings of how well examples of those styles match participants' internal concepts of art. Hypothesis 1 was not supported. In fact, the opposite occurred. Providing contextual information led to participants perceiving examples of the various styles of art as matching less well with their internal standards than when no contextual information was presented. There are a number of possible explanations for this finding.

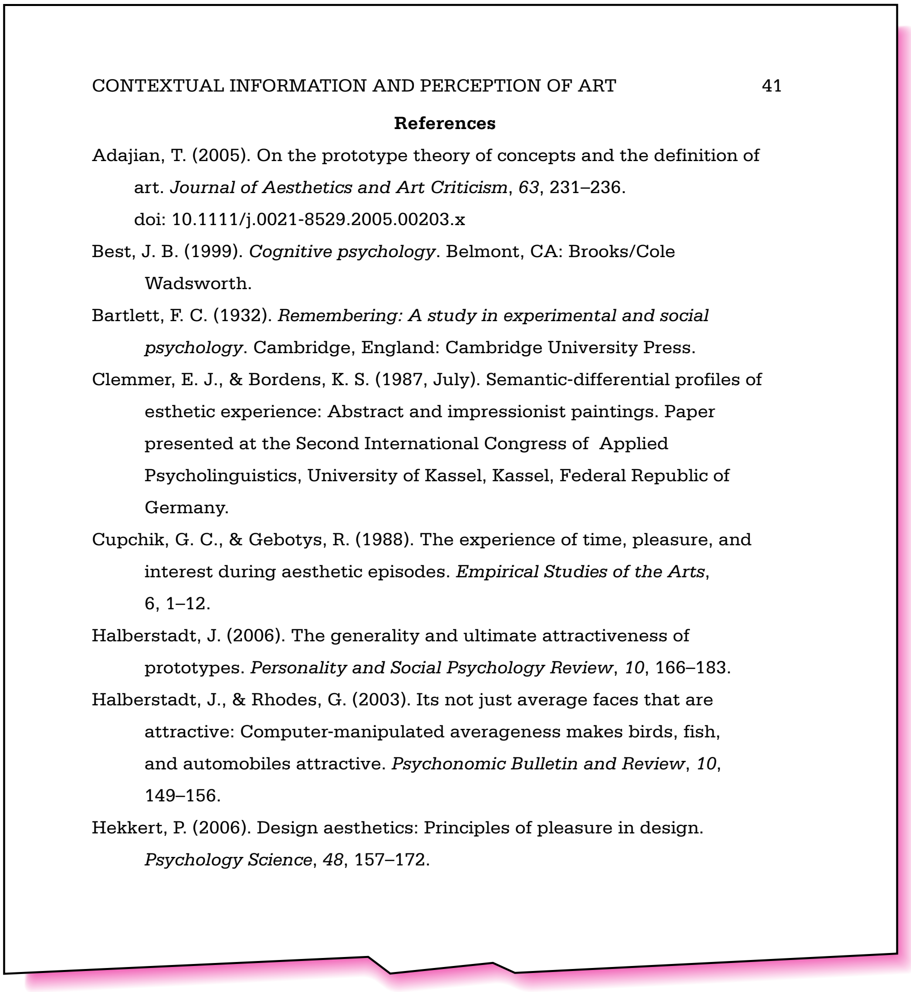
First, initial classification of art is an automatic process (Leder et al., 2004). Providing contextual information may bring this early stage of processing of art under more conscious control, making the issue of just what defines art more consciously available to participants. This greater conscious processing may have led participants to be more critical of the artworks and a corresponding tendency to distance the artworks from their internal prototypes of what constitutes art. Second, there is evidence that enhancing abstract thought is related to individuals willingness to include unconventional styles of art in their internal concept of art than enhancing concrete thought (Schimmel & Förster, 2008). Providing concrete, contextual information on the different styles of art may have encouraged participants to think more concretely about art and thus may have caused them to perceive the examples of art judged as more distant from their internal concept of art than if no contextual information was presented. In fact, in the no contextual information condition, various definitions of art were presented which were less concrete and may have enhanced abstract thought about art resulting in greater acceptance of less conventional art styles. Third, like many category-based judgments, individuals have prototypes representing best-fitting examples of a category. People tend to show a greater preference for exemplars that closely match their internal prototypes. Art is a natural category that may have fuzzy and fluid boundaries defining what constitutes art. Providing contextual information may have caused the prototypes to become more rigid and better defined, making it more likely that particular exemplars would be viewed as less representative of a category than in the no context condition.

FIGURE 16-8 First page of a discussion section. The discussion section begins immediately after the end of the results section, not necessarily on a new page.

The Reference Section

The **reference section** provides a list of the bibliographical references cited in the report. You must list in the reference section all articles, books, or other sources (e.g., conference presentations) that you cited in the body of your paper. Conversely, you *must* cite in your paper any references listed in the reference section. If you read hundreds of papers but only cited three of them, your reference section should contain only the three papers actually cited. Start your reference section on a new page. Type the word "References" (not in boldface) on the line after your running head and page number. Figure 16-9 shows how you format a reference section in APA style.

Use a hanging indent when typing each APA-style reference, as shown in Figure 16-9. Your word processor should allow you to do this easily. For a journal reference, italicize the title of the journal, volume number, and associated punctuation marks. Do not italicize the page numbers, however. Remember to leave a single space after all punctuation marks in a reference. Also include the *DOI system number* (digital object identifier) if provided. You can find the DOI number



CONTEXTUAL INFORMATION AND PERCEPTION OF ART 41

References

Adajian, T. (2005). On the prototype theory of concepts and the definition of art. *Journal of Aesthetics and Art Criticism*, *63*, 231–236. doi: 10.1111/j.0021-8529.2005.00203.x

Best, J. B. (1999). *Cognitive psychology*. Belmont, CA: Brooks/Cole Wadsworth.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge, England: Cambridge University Press.

Clemmer, E. J., & Bordens, K. S. (1987, July). Semantic-differential profiles of esthetic experience: Abstract and impressionist paintings. Paper presented at the Second International Congress of Applied Psycholinguistics, University of Kassel, Kassel, Federal Republic of Germany.

Cupchik, G. C., & Gebotys, R. (1988). The experience of time, pleasure, and interest during aesthetic episodes. *Empirical Studies of the Arts*, 6, 1–12.

Halberstadt, J. (2006). The generality and ultimate attractiveness of prototypes. *Personality and Social Psychology Review*, *10*, 166–183.

Halberstadt, J., & Rhodes, G. (2003). Its not just average faces that are attractive: Computer-manipulated averageness makes birds, fish, and automobiles attractive. *Psychonomic Bulletin and Review*, *10*, 149–156.

Hekkert, P. (2006). Design aesthetics: Principles of pleasure in design. *Psychology Science*, *48*, 157–172.

FIGURE 16-9 First page of a reference section. Start the reference section on a new page.

with the bibliographic information provided in a journal article (usually at the top or bottom of the first page of an article) or in its listing in an electronic database. A specific reference would look like this:

> Schmiege, S. J., Broaddus, M. R., Levin, M., & Bryan, A. D. (2009). Randomized trial of group interventions to reduce HIV/STD risk and change theoretical mediators among detained adolescents. *Journal of Consulting and Clinical Psychology*, *77*, 38–50. doi: 10.1037/a0014513

If the journal is one that starts each issue on page 1 (most journals in psychology do not, but if in doubt you should check), include the issue number in parentheses after the volume number with the issue number not italicized (e.g., . . . *36*(1)). Do not include the issue number for articles appearing in journals that number pages continuously across issues (i.e., issue 1 ends on page 243 and issue 2 begins on page 244).

Table 16-3 provides examples of reference formats for the more commonly cited reference types. If you encounter a reference of a type not covered in the table, consult the APA (2010) publication manual.

For some reason students seem to have a great deal of trouble formatting references properly. We have seen a myriad of errors in reference list entries. With a little attention to detail, you can avoid many of the errors that we commonly see. The following example embodies several of these errors. Compare the reference that follows to the one above and see if you can identify the errors.

> Hardy, Charlie, L., and Van Vugt, Mark. (2006). Nice Guys Finish First: The Competitive Altruism Hypothesis. *Personality and Social Psychology Bulletin*, *32*(10), 1402–1413.

See how many of the following errors that you found:

1. You do not provide authors' full first names (Charlie and Mark). Use the initials only.
2. The "&" sign is used before the last author's name, not the word *and*.
3. Only the first word of the article title is capitalized. Subsequent words use lowercase font (except for the word after the colon).
4. You do not provide the issue number after the volume number of the journal (the [10] shown above) unless the journal is one of the rare ones that starts each issue on page 1.

You list entries in your reference section alphabetically according to the last name of the first author. If there are two articles by the same author from different years, list them in order from oldest to newest, for example, Smith (2005) before Smith (2007). If you have two (or more) references by the same author published in the same year, list them alphabetically according to the title (excluding prepositions such as "a" or "the"). Place lowercase letters immediately after each date, which will correspond to letters assigned to in-text citations for these entries. For example:

> Smith, A. B. (2006a). Control mechanisms in . . .
>
> Smith, A. B. (2006b). A replication of . . .

TABLE 16-3 **Format for Common Reference Sources**

JOURNAL ARTICLE
Print or Electronic Source: DOI Available
Schmiege, S. J., Broaddus, M. R., Levin, M., & Bryan, A. D. (2009). Randomized trial of group interventions to reduce HIV/STD risk and change theoretical mediators among detained adolescents. *Journal of Consulting and Clinical Psychology, 77,* 38–50. doi: 10.1037/a0014513
Print Source: DOI Not Available
Sciangula, A., &. Morry, M. M. (2009). Self-esteem and perceived regard: How I see myself affects my relationship satisfaction. *Journal of Social Psychology, 149,* 143–158.
Electronic Source: DOI Not Available
Stepanova, E. V., & Strube, M. J. (2009). Making of a face: Role of facial physiognomy, skin tone, and color presentation mode in evaluations of racial typicality. *Journal of Social Psychology, 149,* 66–81. Retrieved from http://www.heldref.org/pubs/soc/about.html
REPORT FROM A PRIVATE ORGANIZATION FROM ITS WEB SITE
Alan Guttmacher Institute (2003, September/October). *Services for men at publicly funded family planning agencies, 1998–1999*. Retrieved from http://www.agi-usa.org/pubs/journals/3520203.html
BOOK
Print Version
Lifton, R. J. (1986). *The Nazi doctors: Medical killing and the psychology of genocide*.New York: Basic Books. Tabachnick, B. G., & Fidell, L. S. (1989). *Using multivariate statistics* (2nd ed.). New York: Harper & Row.
Electronic Version of a Print Book
Baldwin, J. M. (1905). *Mental development of the child and the race* (3rd ed.) [Mead Project Version]. Retrieved from http://www.brocku.ca/MeadProject/Baldwin/Baldwin_1906/Baldwin_1906_toc.html
Article in an Edited Book
Austin, W. G. (1986). Justice in intergroup conflict. In S. Worchel & W. G. Austin (Eds.), *Psychology of intergroup relations* (pp. 153–176). Chicago: Nelson-Hall.
PAPER OR POSTER SESSION AT A CONFERENCE
Seeley, W. P. (2008, August). Effects of interpretation of emotional and energetic cost in picture perception. Paper presented at the Twentieth Biennial Congress of the International Association of Empirical Aesthetics, Chicago, IL.

One-author entries come before multiple-author entries in which the first author is the same as the single author. So, for example,

> Adams, J. K. (2004) . . .
>
> Adams, J. K., & Smith, B. D. (2001) . . .

Entries with more than two authors are alphabetized by the last name of the first author, then by the second author, and then by the third author and so on. For example,

> Adams, J. K., Charles, S. L., & Smith, B. D. (2003) . . .
>
> Adams, J. K., Smith, B. D., & Charles, S. L. (2003) . . .

There are other rules for ordering reference entries. For more information see the APA manual (2010, pp. 181–183).

It is increasingly common to find resources published in an electronic medium such as the Internet. For a reference from a journal you obtained online for which a DOI is provided (e.g., through *PsycARTICLES* or EBSCO), use the same format you would had you found the same article in a printed journal (see the first example in Table 16-3). If, on the other hand, you use an online version of an article and no DOI is available, provide the home Web site address of the journal (APA, 2010, p.199) as shown in the following example:

> Stepanova, E. V., & Strube, M. J. (2009). Making of a face: Role of facial physiognomy, skin tone, and color presentation mode in evaluations of racial typicality. *Journal of Social Psychology, 149,* 66–81. Retrieved from http://www.heldref.org/pubs/soc/about.html

Notice a few things about this format. First, everything is the same as for a print journal reference item. Second, provide the home Web page for the journal and *not* the Web address of the database you used to find the article. You do this because the information included in a database may change frequently, and presumably, the journal's Web site will be more constant (APA, 2010). Do not put a period at the end of the Web address. Finally, do not provide the date on which you retrieved the document.

Footnotes

APA writing style specifies two types of footnote. A *content footnote* clarifies a point made in the text of the paper, or to provide additional details that would detract from the flow of your discussion at that point. The other type of footnote, the *copyright permission footnote,* acknowledges the source of copyrighted quoted material, figures, or tables.

In the body of your paper, you number content and copyright footnotes consecutively using Arabic numerals. At the point in the text at which the reader should consult the footnote, simply place a superscript number (beginning with 1 for the first footnote), as in the following example:

> The trial, used previously (Horowitz & Bordens, 1988), consisted of a 4-hour audiotape.^{1}

There are two options for footnote placement in your manuscript. You may place each footnote on the bottom of the page on which the noted material appears. You can do this easily by using the footnote function of your word processor. Alternatively you may place the footnotes on a separate page after your references, entitled "Footnotes."

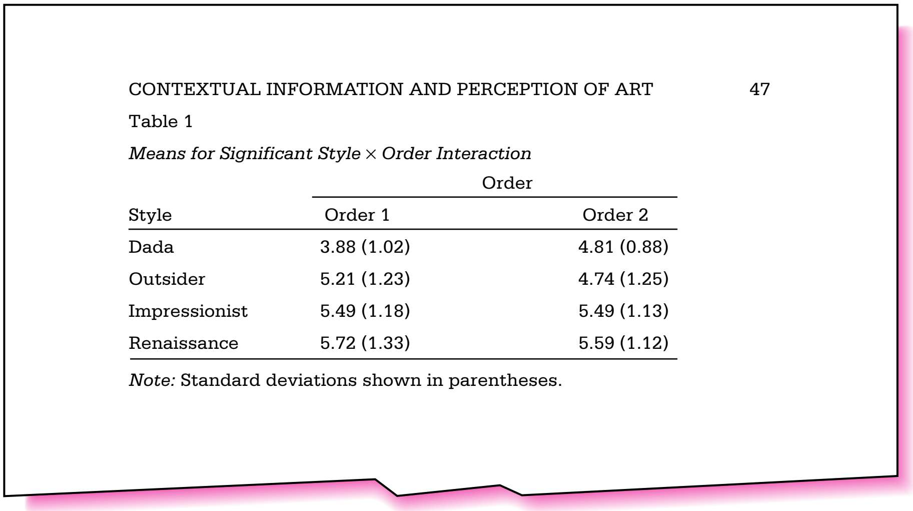
QUESTIONS TO PONDER

1. Where do you begin the reference section in an APA-style manuscript?
2. What information do you include in an APA-style reference and how is a reference entry formatted?
3. How are footnotes used in an APA-style manuscript and where are they placed?

Tables

You use tables to present complex information that you cannot easily summarize in the body of your paper. For example, they can illustrate the design of your study or present summary data (e.g., tables of means and standard errors, correlation matrices). Tables are somewhat time consuming to make and expensive to reproduce. Use a table only when you cannot fully describe information in the text of your paper.

Tables prepared according to APA specifications include a title, a number, headings, a body, and, if necessary, notes. Create a separate page for each table, which should appear as shown in Figure 16-10. Place the title and number of the table at the top of the page, as illustrated. The headings of your table should clearly tell your reader what information is included in your table. In the body of the table, include



CONTEXTUAL INFORMATION AND PERCEPTION OF ART 47

Table 1

Means for Significant Style × Order Interaction

	Order	
Style	Order 1	Order 2
Dada	3.88 (1.02)	4.81 (0.88)
Outsider	5.21 (1.23)	4.74 (1.25)
Impressionist	5.49 (1.18)	5.49 (1.13)
Renaissance	5.72 (1.33)	5.59 (1.12)

Note: Standard deviations shown in parentheses.

FIGURE 16-10 Sample APA-style table.

the information that you want your reader to see. Finally, as shown in Figure 16-10, use notes to explain the meaning of symbols in the table or to provide information not included in the table itself. Place pages with tables after your reference section.

Figures

Use figures in your paper to provide graphic illustrations of complex material or relationships that cannot be adequately described in text. Although figures appear most often in the results section of your paper, they also can appear in any other section. For example, you could use a figure in your method section to illustrate the materials used in your study or in the introduction to show an important theoretical relationship. Because figures are difficult to prepare and are expensive to reproduce in journals, use them sparingly.

Graphs, drawings, and photographs are three commonly used types of figures. Use graphs to illustrate complex relationships among variables. Use drawings and photographs to illustrate equipment, materials, or stimuli you used in your study. You might use photographs to convey aspects of results that you cannot adequately describe in the text of your paper. For example, articles about physiological psychology may present photographs of histological sections to show the location of a lesion or stimulating electrode in the brain.

A simple rule to follow when preparing graphs is to make them simple and accurate. A visually confusing graph adds little to your paper. An improperly drawn graph can confuse your reader or make small, albeit statistically significant, effects look unnaturally large. Carefully plan and draw your graphs.

Figure 16-11 illustrates how the scaling of an axis on a graph can influence the reader's impression of the size of an effect. Participants scanned a one-page document for typographical errors under dim, medium, and bright lighting. The document contained five errors. Panel (a) displays the average number of errors found (vertical axis) as a function of lighting intensity (horizontal axis). A cursory look at this graph conveys the impression that lighting intensity had a large influence on the number of errors detected. However, the examination of the scale along the vertical axis reveals that the entire range of values varies only between 3.6 and 4 errors. The graph shown in panel (b) shows a more accurate impression of the influence of lighting intensity

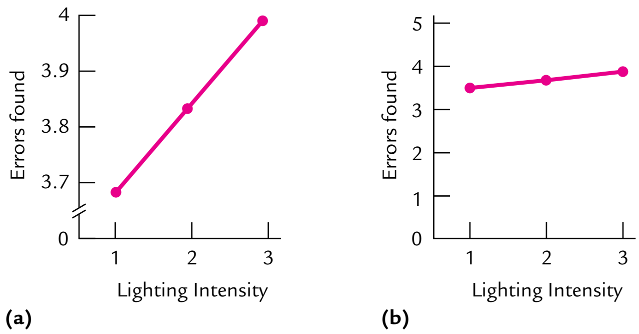


FIGURE 16-11
(a) Graph that exaggerates size of a relationship;
(b) graph that accurately depicts a relationship.

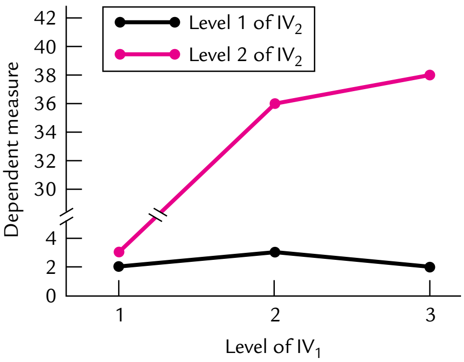


FIGURE 16-12 Graph showing a broken axis.

on the number of errors detected. Here the scale of errors presented along the horizontal axis covers the entire range of possible values (0 to 5). The small differences among means are represented fairly, and your reader is not misled.

In some cases, you may find it necessary to break an axis to fit all your data on a graph. For example, if values of the dependent variable can range from 0 to 50 but participants used only numbers between 0 and 3 and between 32 and 40, you may want to show the data on a graph similar to the one shown in Figure 16-12. Notice that the *y*-axis is broken by slash marks. This shows your reader that a range of values along the *y*-axis has been skipped over. You should also break any lines within the graph that cross the broken part of an axis, as shown in the figure.

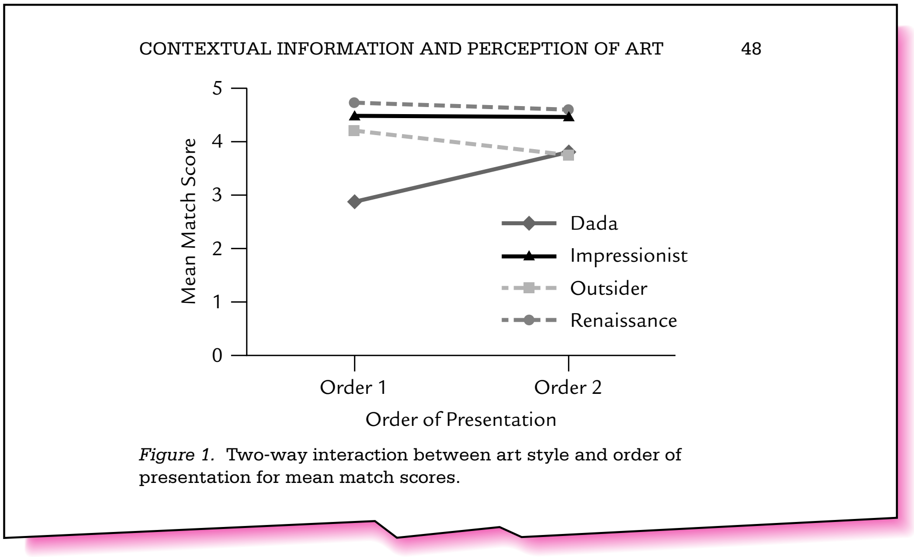
Some graphs require a *legend* to identify the meanings of the symbols or line styles used in the graph. The legend appears within the figure itself. Look again at Figure 16-12, which shows how an APA-style figure page looks. The legend at the top of the graph tells the reader which line represents which level of art style.

You must also include a caption for your figure. The caption provides the title for your figure along with any other necessary explanatory information (e.g., the source of a figure, information about what specific numbers mean). Type the caption immediately below your figure using the format shown in Figure 16-13.

If you have more than one figure, each goes on its own page. Each page is included as part of the manuscript and must have your running head and a page number on it. Position the pages with your figures after any pages with tables that you have included.

QUESTIONS TO PONDER

1. When are tables used in an APA-style manuscript?
2. How are the tables used in an APA-style manuscript formatted?
3. When do you use figures in an APA-style manuscript?
4. How is a page containing a figure set up and what is included on a figure page?



CONTEXTUAL INFORMATION AND PERCEPTION OF ART 48

Figure 1. Two-way interaction between art style and order of presentation for mean match scores.

FIGURE 16-13 Sample APA-style figure.

ELEMENTS OF APA STYLE

In addition to knowing about the different sections of an APA-style manuscript and what goes in them, you must also know some things about conventions used in APA style. These include citing references in your paper, citing quoted material, using numbers in your paper, and avoiding biased language. We discuss these conventions next.

Citing References in Your Report

In some writing styles, you indicate citations in the body of your paper with footnotes. In APA style, however, citations are made by providing the name(s) of the author(s), the publication date of the source, and (when needed) specific pages within the source.

The format of a citation within the text of your manuscript depends on how you choose to write a sentence that includes the citation. If the citation is included as an integral part of the sentence, you give the last name(s) of the author(s) and, in parentheses, the year in which the work was published. Here is an example with two authors:

> According to Smith and Jones (2009), memory for meaningful information is much better than memory for meaningless information.

If the citation is "tacked on" to the sentence, enclose the entire citation in parentheses, as follows:

> Memory for meaningful information tends to be much better than memory for meaningless information (Smith & Jones, 2009).

The two examples just provided are *multiple-author citations*. Notice that when you cite the names of two authors in the sentence itself (the first example), you use the word *and* to connect the authors' names. When the names are used in parentheses (as in the second example), you use an ampersand (&) to connect them.

When you have more than two authors, the citation takes the following form:

> According to Smith, Jones, and Harris (2009), memory for meaningful information is better than memory for meaningless information.

or

> Memory for meaningful information is better than memory for meaningless information (Smith, Jones, & Harris, 2009).

When a citation is from a source with only two authors, *always* provide the names of both authors each time you cite the source. However, if you must cite an article with three or more authors several times in your paper, writing each name repeatedly becomes tedious. In this case, provide all of the authors' names the first time you cite the source (e.g., Smith, Jones, & Harris). Thereafter, you can provide only the first author's name followed by "et al." (Latin for "and others"). Here is an example:

> Memory for meaningful information is much better than memory for meaningless information (Smith, Jones, Harris, Baker, & Thomas, 2009) . . . Smith et al. also point out that . . .

If you cite a source with more than six authors, use the "et al." device for the first citation and all subsequent citations. In the reference section, you would provide the last names and initials for the first six authors and et al. for the remaining authors.

What happens if you have two citations from the same authors in different orders from the same year (e.g., Smith, Jones, Harris, & Baker, 2008, and Smith, Harris, Jones, & Baker, 2008)? Using the et al. device would be confusing because your reader would not know to which source you are referring. In this case, you provide as many author names as needed to distinguish between the two sources (e.g., Smith, Jones, et al., 2008, and Smith, Harris, et al., 2008).

Finally, several special in-text citations are sometimes used. These are summarized in Table 16-4.

Citing Quoted Material

Whenever you directly quote a source, you must indicate that the material was obtained from another source. You must include the author's name, year of publication, and page or pages where you found the quoted material. For electronic sources without page numbers, provide the paragraph number (e.g., para. 2). If the electronic source has headings, provide the section title (e.g., Results section, para. 3). For shorter quotes (40 words or less) include the quoted material in your paragraph and enclose the quoted material in quotation marks. As with reference citations, the form to follow depends on the sentence structure. For example,

> Although research does suggest that television has the potential to aid in the socialization of children, there is still reason to be cautious. In fact, according to Liebert, Sprafkin, and Davidson (1982, p. 209), "Although most studies

TABLE 16-4 APA-Style, In-Text Citations

APPLICATION	CITATION FORMAT
Single Author	
AUTHOR NAMED IN SENTENCE	
One article:	Jones (2010)
Two articles (same year):	Jones (2010a, 2010b)
Two articles (different years):	Jones (2009, 2010)
Personal communication:	Smith (personal communication, July 15, 2010)[a]
AUTHOR NAMED IN PARENTHESES	
One article:	(Jones, 2010)
Two articles (same year):	(Jones, 2010a, 2010b)
Two articles (different years):	(Jones, 2009, 2010)
Personal communication:	(Smith, personal communication, July 15, 2010)[a]
Multiple Authors[b]	
AUTHORS NAMED IN SENTENCE	
Two authors:	Smith and Jones (2010)
More than two authors:	Smith, Jones, and Key (2010)
AUTHORS NAMED IN PARENTHESES	
Two authors:	(Smith & Jones, 2010)
More than two authors:	(Smith, Jones, & Key, 2010)
Multiple citation for same idea:	(Harris, 2008; Jones, 2010; Smith & Jones, 2006)
Special Citations	
LEGAL CITATIONS	
Litigants named in sentence:	*Ballew v. Georgia* (1976)
Litigants named in parentheses	(*Ballew v. Georgia*, 1976)

[a] Personal communication items are *not* placed in the reference list.

[b] Format for multiple articles follows that for single authors.

SOURCE: Information compiled from American Psychological Association, 2010.

> suggest that prosocial television can have desired effects, our ability to magnify these effects and minimize undesirable ones is in its infancy."

or

> Even though research does suggest that television has the potential to aid in the socialization of children, "our ability to magnify these effects and minimize undesirable ones is in its infancy" (Liebert, Sprafkin, & Davidson, 1982, p. 209).

or

> Even though research does suggest that television has the potential to aid in the socialization of children, Liebert, Sprafkin, and Davidson (2006) say that "our ability to magnify these effects and minimize undesirable ones is in its infancy" (p. 209).

For longer quotes (40 or more words), set the quoted material off in a block paragraph format without quotation marks. Type the entire block quote about a half inch from the left margin. The first line of the first paragraph is not indented further. The first lines of subsequent paragraphs of the block quote are indented an additional half inch. Provide appropriate citation information at the end of the paragraph if you did not provide it in the sentence that introduced the quoted material. For example:

> For example, John Watson once famously said
>
> > Give me a dozen healthy infants, well formed, and my own specified world to bring them up in and I'll take any one at random and train him to become any type of specialist I might select—doctor, artist, lawyer, merchant-chief and, yes, even beggar-man and thief, regardless of his talents, penchants, tendencies, abilities, vocations, and race of his ancestors (Watson, 1930, p. 104).

These examples cited a direct quotation. Even when you simply paraphrase someone's ideas or words, you still cite the source (although you do not provide page numbers for paraphrased ideas). We discuss this issue in detail later in this chapter in the section on plagiarism.

QUESTIONS TO PONDER

1. What are the general rules for in-text citations?
2. How do you cite quoted material in your paper?

Using Numbers in the Text

As a rule, you spell out numbers lower than 10 (e.g., five), and express numbers 10 and above numerically (e.g., 23). A few exceptions to this rule are as follows (APA, 2010). Use numerals in these cases:

1. Use numerals when a number immediately precedes a unit of measurement (e.g., a 5-mg dose).
2. Use numerals when representing statistical or mathematical functions, percentages, ratios, percentiles, and quartiles (e.g., multiplied by 4; the 2nd percentile).

3. Use numerals when representing time, dates, ages, sample or population sizes, specific numbers of subjects in an experiment, scores and points on a scale, sums of money, and numerals as numerals (e.g., the longitudinal study took 4 years; numbers on the scale ranged from 0 to 10). An exception to this rule is to write out numbers that refer to approximated time periods (e.g., approximately sixteen years).
4. Use numerals when discussing a specific place in an ordered series, parts of books, and each number in a list of four or more numbers (e.g., Grade 2; Table 1).
5. Use numerals for all numbers that appear in the abstract of a paper or in graphical displays.

Express numbers in words in the following cases (APA, 2010):

1. Any number that begins a sentence, title, or heading is written out (e.g., Twenty subjects were assigned; Three subjects were dropped from the analysis, leaving 58 subjects). However, consider rewriting sentences or headings that begin with numbers so that they do not begin with a number.
2. Write out numbers representing common fractions (e.g., two-thirds of the class).
3. Write out numbers for universally accepted usages (e.g., Five Books of Moses, the Twelve Apostles).

Additional rules for using numbers in APA style are too lengthy to cover here. See the APA publication manual (2010, pp. 111–115).

Avoiding Biased Language

Biased language occurs when you inadvertently use language that presupposes that one group is preferred over another or that might be offensive to a group of people. For example, by using the male pronoun *he* to express ideas generically, we are falling into the trap of biased writing. The APA publication manual gives three guidelines to follow to avoid biased language.

Guideline 1 refers to providing descriptions at the appropriate level of specificity. For example, when you describe ethnic groups, avoid using general terms such as *Asian* or *Hispanic*. Instead, be specific. If your participants were Chinese or Puerto Rican, use those more specific terms. When you refer to sexual orientation, rather than using the term "gay" to refer to men and women, use "gay men" or "lesbian women" instead (APA, 2010). There may also be some confusion over when to use the term "gender" or "sex." The term "gender" is used when referring to men and women in a social or cultural context. The term "sex" is a biological term and is used when referring to biological differences between men and women (APA, 2010). There are two general rules to follow with respect to describing groups. First, when in doubt always use a more specific term; second, use group descriptors only when they are relevant (APA, 2010).

Guideline 2 deals with the issue of "labeling." You should be sensitive to the labels you attach to people. A good rule to apply is to avoid labeling groups of people. For example, the label "the elderly" categorizes people as if they were objects

(APA, 2010). The term "elderly people" is preferred (APA, 2010). Also in this category of bias is writing that elevates one's own group above others. For example, the phrase "men and wives" implies that males are the standard by which women are to be judged (APA, 2010). In this case, "men and women" would be better. Finally, be aware of the fact that the order in which you list groups may imply superiority of one group over another. For example, always mentioning White participants before Black participants may imply a superiority of Whites over Blacks. To avoid this, mix the order in which you present group identifiers.

Guideline 3 suggests that you refer to individuals who participate in your research in a way that acknowledges their participation. For example, rather than saying, "Participants were run in groups of four," say, "Participants completed the experiment in groups of four." The latter sentence conveys that the participants were active in the experiment, which they were.

Space does not allow us to discuss in detail how to avoid biased language. The APA publication manual (2010, pp. 70–77) provides an extended discussion of the three guidelines just outlined as well as discussions of specific areas of concern (e.g., racial and ethnic identity, gender, sexual preference, and disabilities). You should consult the APA publication manual (2010) for further information on how to avoid biased language. One thing to keep in mind, however, is that avoiding biased language should not come at the expense of precision and accuracy. An attempt to avoid offending anybody can cloud important information. The APA urges that you use good judgment rather than a set of rigid rules concerning what is acceptable writing.

QUESTIONS TO PONDER

1. What are the general rules for using numbers in the text of a manuscript? What are the main exceptions to the general rules for using numbers?
2. What is biased language and why should you avoid using it?
3. What are the three APA guidelines for avoiding biased language?

EXPRESSION, ORGANIZATION, AND STYLE

The previous sections explained the general conventions to follow when writing an APA-style paper. Unfortunately, merely knowing about subsections and citation formats does not guarantee that you will write a quality paper. You also must know how to present your ideas in a clear and organized way.

As a rule, you should write using the active voice in a sentence. That is, sentences should follow a subject–verb–object organization. The two examples that follow show two sentences, one written in passive voice and the other in active voice.

Passive voice: A questionnaire was given to each participant to complete.

Active voice: I gave a questionnaire to each participant to complete.

Using the personal pronouns *I* or *we* is acceptable in APA style and is even preferable because it encourages writing in the active voice. In years past, authors were encouraged to write in the third person. Current APA style allows you to write in the first person.

Perhaps the most common flaws in student papers are poorly expressed ideas, unorganized presentation of ideas, and sloppy presentation style. You can have a good handle on your research topic, procedures, and results but still be unable to communicate them well to your readers. Unclear writing and a disorganized presentation obscure important points. Although this chapter cannot teach you to be a good writer, it can help you avoid some of the common pitfalls. This section points out some of the flaws commonly found in student papers.

Precision and Clarity of Expression

In your writing, you should express your ideas clearly and concisely to your readers. We explore three elements of clear expression in the sections that follow: grammatical correctness, proper word choice, and economy of expression (APA, 2010).

Improper grammar can seriously interfere with the clarity and precision of your manuscript. Unfortunately, students often do not pay close enough attention to the grammatical structure of their sentences. A result of this can be ambiguous sentences. Consider the following example:

> The experimenter recorded how fast each rat ran the maze with a stopwatch.

In this example, the writer's intention was to say that the experimenter used a stopwatch to time the rat's running speed. Instead, you get the image of a rat running the maze while holding a stopwatch! The problem here is that the modifier "with a stopwatch" is misplaced in such a way that it refers to the rat, not to the experimenter. The following rewritten sentence is unambiguous and consequently much clearer:

> Using a stopwatch, the experimenter timed each rat's running speed.

Here, the modifier is properly attached to the experimenter.

Another common grammatical error is disagreement between subject and verb. In the following incorrect sentence, the verb goes with the word "records" and not "one":

> Only one of the subject's *records were* included in the analysis.

This sentence should be corrected as follows:

> Only *one* of the subject's records *was* included in the analysis.

Of course, this chapter cannot explore all the common grammatical errors. But you should make an effort to reduce grammatical errors in your writing. Some good guides that may help you are Crews (1980); Hall (1979); Leggett, Mead, and Charvat (1978); Strunk and White (1979); and Chapter 3 of the APA publication manual (2010). You can also use an online source such as the Purdue University Owl Web site (http://owl.english.purdue.edu/owl/). A grammar checker, included with most word processors, may also help (but is not always correct).

Proper word choice is the second aspect of clear writing. Make sure that you select words that convey your ideas as you intend. For example, in general writing, the words *feel* and *believe* are used interchangeably. In scientific writing, however, they may mean very different things (APA, 2010). Choose the word that most accurately conveys your meaning. If you mean *believe*, then say *believe*.

Unnecessary qualifiers also tend to reduce clarity of expression. Phrases such as "approximately equal" and "particularly strong" are imprecise, and different readers may interpret them differently. Be specific when referring to quantity estimates. Moreover, be careful about using pronouns in place of nouns, especially in long sentences. Following the meaning of a sentence that has many "they's," "it's," and so on is difficult.

Avoid using complex words when a simple word would suffice. For example, don't say *masticate* when you mean *chew,* or *cogitate* when you mean *think* (Hall, 1979, p. 84). In addition, never utilize the word *utilize*—use *use* instead. Constant use of fancy words becomes tedious to read. Base your choice on how well a word conveys your meaning, not on how intelligent you think it makes you sound. If a complex word best conveys your meaning (e.g., *clandestine* implies much more than *secret*), use it. In other cases, use the preferred simple word.

Using jargon can also reduce clarity. *Jargon* refers to the use of technical vocabulary even when that vocabulary is not relevant. Not all readers will know what you mean by these terms, so it is best to avoid them if possible. For example, using the term *voir dire* throughout an article may confuse readers who are not familiar with legal terminology. It would be better to use the term *jury selection*.

There are other flaws that can affect precision and clarity. See the APA manual (pp. 68–70) for a discussion of these flaws.

Economy of Expression

You should express your ideas in a concise, economical manner (APA, 2010). Three major flaws in writing threaten economy of expression: the use of wordiness, redundancy, and unit length (APA, 2010).

Wordiness refers to using more words than necessary to express an idea clearly. Consider the following examples:

Wordy: There were several participants who required additional assistance.

Better: Several participants required additional assistance.

The currently popular phrase "the exact same" is not only grammatically suspect ("exactly the same" would be better); it uses three words when one will do: "identical."

Redundancy occurs when words duplicate a meaning already conveyed by other words. In the following example, the word *past* is redundant: "The participant's past history was examined." *Unit length* concerns the length of sentences within a paragraph. Using too many short sentences creates a choppy, boring style. Using excessively involved, long sentences can be confusing. Vary the length of the sentences within a paragraph to establish and maintain the interest of your reader. Attention to unit length also extends to the length of paragraphs (APA, 2010), a topic we discuss in the next section.

Organization

Whereas clarity of expression relates most closely to the structure of sentences, organization relates to how you organize those sentences into paragraphs and how you weave paragraphs into an entire paper. You can write the most beautiful, grammatically correct sentences yet fail to convey your ideas clearly. A paragraph is more than a collection of grammatically correct sentences. You must weave those sentences into a coherent, unified entity that clearly conveys information to your readers.

Paragraphs can include four types of sentences (Crews, 1980). These are theme sentences, support sentences, limiting sentences, and transitional sentences. The *theme sentence*, which is usually the first sentence in a paragraph, conveys to your reader the topic of the paragraph. *Support sentences* follow the theme sentence and support and elaborate the theme. *Limiting sentences* point out possible limits to the assertion made in the theme sentence. Finally, *transitional sentences* are used to shift smoothly from one idea to another within a paragraph. These four types of sentences should be combined to achieve unity.

Crews (1980) suggests four general rules to help you attain unity within paragraphs:

1. Make only one major point within a paragraph.
2. Make your theme sentence the most general sentence within a paragraph. All subsequent sentences should focus on the theme stated in the theme sentence.
3. Stick to the theme stated at the outset of the paragraph. Unity is disrupted when you stray from the point.
4. Use complete sequences of limiting or supporting sentences. That is, if you have something positive to say about a topic, list all the positive elements before turning to the negative.

The following paragraph exemplifies these points:

> Piaget's view of the development of object permanence was based on rather informal tests with infants. In a typical test, Piaget hid an object, and the infant was required to engage in a visual or manual search for it. From the results of such tests, Piaget concluded that infants do not demonstrate object permanence until they are 6 months old and do not achieve a full understanding that objects continue to exist when out of sight until the end of the sensorimotor stage. However, later research suggests that infants may acquire the concept of object permanence long before Piaget suggested they do.

The first sentence in the paragraph is the theme sentence because it tells the reader that the paragraph is about Piaget's views on object permanence. The subsequent sentences discuss issues relevant to the theme sentence. The next two support the theme sentence, whereas the last one limits it.

In addition to unity, a paragraph should have coherence. Coherence is disrupted when sentences that relate only tangentially to a topic are included in a paragraph (Hall, 1979, p. 193). Two problematic sentences are italicized in the following paragraph. The first disrupts unity, and the second disrupts coherence.

> Methodological problems were apparent in the Jones and Smith (2010) experiment. *Smith was only a graduate student at the time of the study*. The methodological problems stemmed from use of outdated equipment. *Some of the equipment was so old that it could hardly be kept working*. This equipment did not have the sensitivity to record accurately the subtle changes in the subjects' behavior.

The first italicized sentence is irrelevant to the topic introduced at the beginning of the paragraph. The second italicized sentence, although tangentially related, is not necessary. It serves only to break up the important points in the previous and subsequent sentences (points that you should link directly). The two italicized sentences together add little to the discussion and disrupt the unity of the paragraph.

Without these two unnecessary statements, the paragraph still conveys the important idea that the Jones and Smith study was flawed because of archaic equipment:

> Methodological problems were apparent in the Jones and Smith (2010) experiment. The methodological problems stemmed from use of outdated equipment. This equipment did not have the sensitivity to record accurately the subtle changes in the subjects' behavior.

Paragraphs can become confusing when they become too long. Overly long paragraphs are a common error we have encountered in student writing. If a paragraph runs longer than one double-spaced manuscript page, look for places to break the paragraph into shorter ones (APA, 2010). Breaking long paragraphs into shorter ones provides pauses for your readers and is preferable to a single, overly long paragraph. Weave together paragraphs to create a unified and consistent narrative that will hold your readers' interest. Avoid confusing your readers with too much information packed into too little space.

On a similar note, you also should avoid using a sequence of overly short paragraphs. Overusing one- or two-sentence paragraphs breaks up the flow of your writing and makes it abrupt and difficult to read (APA, 2010). If you find you have many short paragraphs, see if you can reorganize your ideas to create fewer but longer paragraphs.

You should organize each section of your paper into units. For example, you might organize your introduction into three subsections (even if they are not labeled as such). In the first subsection, you introduce your topic (using perhaps two paragraphs). In the second subsection, you might review relevant literature (five paragraphs). Finally, in the third subsection, you might summarize the research you reviewed and state your hypotheses (three paragraphs). You should organize the information in your method, results, and discussion sections in a similar fashion.

The best way to avoid disorganization is to make an outline of your paper before you begin writing and then stick to the outline. Indicate the main sections (introduction, method, etc.), and identify the subtopics to be handled within the major sections.

Style

Your final paper is a reflection of you as well as of your work. A paper can be well organized and clearly written yet still make a negative impression on your reader because of sloppy presentation. Frequent misspellings, misused words, and typographical errors detract significantly from your paper. Work to eliminate them.

You can best avoid misspellings by using a dictionary (we recommend *Webster's 11th edition*, *Webster's Third New International edition*, or the Merriam-Webster Online Dictionary [http://www.m-w.com/dictionary.htm]). If you have difficulty with spelling (as many of us do), have your paper read by a good speller. The spelling checker that comes with your word processing program also can help identify some spelling errors. However, it will not catch errors involving "sound-alike" words (such as using "witch" when you meant "which") or the wrong grammatical form (e.g., typing "our" when you meant "your"). In addition, it may not recognize certain technical terms even when these are spelled correctly. For those you will need to keep your dictionary handy.

Misused words also detract from your paper. For example, "affect" and "effect" are commonly confused. As a verb, "affect" means "to act on," and "effect" means "to bring about" (see Table 16-5 for a list of commonly misused words). Avoiding these errors involves acquiring a good vocabulary. If you have trouble in this area, have your paper read by someone who has a good vocabulary. In addition, your word processor probably includes a facility that flags possible grammatical errors and suggests alternative constructions. Although these are not always accurate, they do help you to spot potential trouble spots.

Frequent typographical errors (including crossed-out words, penciled-in words, and mistyping) also detract from your work. Careful proofreading is essential. Correct any errors that you find before you submit your work. You should make these corrections on your word processor, not pencil them in on the manuscript. Today's word processors are invaluable for making corrections. They allow you to insert, delete, or move words or even whole paragraphs quickly and easily. You can fix errors, make stylistic changes, or improve organization without retyping sections of the manuscript.

QUESTIONS TO PONDER

1. Why are precision and clarity of expression, organization, and style so important to consider when preparing a manuscript?
2. What factors contribute to or detract from precision and clarity of expression?
3. What factors contribute to or detract from good orgnization?
4. What can you do to ensure that your paper has proper style?

Making It Work

A goal that you should strive to achieve is to produce a well-written report of your results that is clear, organized, and visually pleasing. Most writers, even professionals, cannot produce a "finished product" after only a single writing. The best way to approach writing a paper (especially in the early stages of your writing career) is to prepare a rough draft and then make careful revisions. If your university or college has a writing center (or other similar resource), you can have someone there read the draft of your paper. This person can point out flaws and give you ideas about correcting those flaws.

TABLE 16-5 Commonly Misused Words

WORDS	TRUE MEANINGS AND COMMENTS
affect/effect	*affect:* to influence *effect:* the result of; to implement
accept/except	*accept:* to take willingly *except:* excluding; to exclude
among/between	*among:* used when you refer to more than two *between:* used when you refer to only two
amount/number	*amount:* refers to quantity *number:* refers to countable elements
analysis/analyses	*analysis:* singular form *analyses:* plural form
cite/site	*cite:* make reference to *site:* location
datum/data	*datum:* singular form *data:* plural form
every one/everyone	*every one:* each one *everyone:* everybody
few/little	*few:* refers to number *little:* refers to amount
its/it's	*its:* possessive pronoun *it's:* contraction of "it is"
many/much	*many:* refers to countable elements *much:* refers to quantity
principle/principal	*principle:* strongly held belief *principal:* foremost
than/then	*than:* conjunction used when making a comparison *then:* refers to the past in time
that/which	*that:* used to specify a crucial aspect of something: "the study that was conducted by Smith (1984)" *which:* used to offer a qualification that is not crucial to something: "the study, which was published in 1984" (*which* is always preceded by a comma; *that* takes no comma)
there/their/they're	*there:* refers to a place *their:* possessive pronoun *they're:* contraction of "they are"
whose/who's	*whose:* the possessive of "who" *who's:* contraction of "who is"
your/you're	*your:* possessive pronoun *you're:* contraction of "you are"

SOURCE: Compiled from Crews, 1980; Hall, 1979; Leggett, Mead, & Charvat, 1978; and Strunk and White, 1979.

During the revision process, look for three major things. First, read through your paper, paragraph by paragraph, and check for unity, coherence, and proper word usage. Second, read your paper for organization between paragraphs and sections. Finally, carefully comb your paper for typographical errors, misused words, and other stylistic errors. Only after you have completed several cycles of writing and revising should you submit your paper to your instructor or to a journal.

Avoiding Plagiarism and Lazy Writing

Reference citations must be included in your paper to give credit to another person or persons who have published or presented ideas. If you use someone else's words or ideas without proper citation, you are guilty of **plagiarism**, which is at best unethical and at worst illegal. Penalties for plagiarism can range from a failing grade on an assignment to civil litigation (if the plagiarized work is published).

A broad rule of thumb to avoid plagiarism is to provide a citation whenever another person's work influenced your thinking. Of course, this means citing more than direct quotations and paraphrases of someone else's writing. If an idea you present in your paper is not originally yours, then you must cite the source.

A writing deficiency closely related to plagiarism is lazy writing (Rosnow & Rosnow, 1986). In **lazy writing**, an individual simply lifts paragraph after paragraph out of one or more sources and presents them as a paper. The difference between plagiarism and lazy writing is that in lazy writing, the individual properly cites the source of the material.

Although lazy writing is technically not plagiarism, few instructors will accept a paper that relies heavily on quoted material. Keep the following rules of thumb in mind when writing a paper:

1. Always properly cite the source of words and ideas that are not your own.
2. Always paraphrase information from another source and provide a proper citation.
3. Enclose directly quoted material in quotation marks or set longer passages off in a block paragraph style and provide the proper citation, which includes the page number(s) where the material can be found in the original source.
4. Use quoted material sparingly and *only* to support something you have written *in your own words*.
5. Make sure any written assignment that you turn in is written in your own words. Never turn in a paper that consists of large amounts of material taken from other sources with little of your own writing. This is true even if you made some minor, cosmetic changes to the original material and properly cited the original source.

For further information on lazy writing, go to the Web site supporting this text. There you will find extended examples of plagiarism and lazy writing as well as advice on how to avoid these two writing flaws. You also will find links to a number of Web sites that deal with plagiarism.

QUESTIONS TO PONDER

1. What are plagiarism and lazy writing?
2. How can you avoid plagiarism and lazy writing?

TELLING THE WORLD ABOUT YOUR RESULTS

If you decide to pursue a career as a psychologist, you will probably need to submit your research to a journal for publication. Once you have prepared your APA-style paper, you must make a few decisions about how to disseminate your results, or make your results known, to the scientific community. You have several options, which are not necessarily mutually exclusive. You may present your results at a local, regional, or national convention (e.g., the Annual Meeting of the Midwestern Psychological Association), an option that we discuss below. You also may decide to have your results published in a scientific journal, or you could publish your results on a Web site on the Internet.

Publishing Your Results

Before submitting your paper to a scientific journal for possible publication, you have to make a couple of preliminary decisions. One decision is where to send your paper. If your paper has a highly specific focus (such as treating abnormal behavior), you should consider sending your paper to a specialized journal (for this example, perhaps the *Journal of Abnormal Psychology*). If your paper is more broadly focused, you might send it to a less specialized journal. For example, if you conducted a study of the attributions made by schizophrenic patients, you might send your paper to the *Journal of Personality and Social Psychology*. Refer to Chapter 3 for a list of some of the major psychological journals.

A second decision is whether to send your paper to a refereed or nonrefereed journal. As stated in Chapter 3, papers sent to a refereed journal are reviewed before publication, but those sent to a nonrefereed journal are not reviewed. Select a refereed journal because work published in a refereed journal usually receives more serious attention by the scientific community.

The Publication Process After you submit your paper to a refereed journal, it goes through a standard procedure for review and publication. First, your paper is sent out for review. After the initial review, you receive a decision about acceptance or rejection. The editor of the journal may unconditionally accept or reject your paper. In many cases, however, papers are given a *conditinal acceptance*. The editor then asks you to revise and resubmit your paper. After resubmission, the paper may be sent out for another review if the revisions are extensive.

Once accepted, your paper goes to a *copy editor*, who makes sure that the paper conforms to the style and requirements of the journal. In most cases, the copyedited manuscript will be returned to you so that you can review it. At this point, you may still make limited changes to the manuscript. Once you have reviewed the manuscript and corrected any errors, you then return the copyedited manuscript to the editor. Many journals allow you to return your manuscript electronically to the publisher for production.

After the publisher has typeset your manuscript, you receive *proofs*, which are copies of your paper as it will appear in the journal. You read these proofs to be sure that they agree with your manuscript and that there are no errors. Errors may be of two types: printer errors and author errors. A printer error occurs when the text in the proofs does not match the "submitted" manuscript. You would indicate these errors and identify them as printer errors. Author errors are things you did not catch in the previous rounds of editing. For example, you may have discovered that you reported a statistic incorrectly. You can fix these at the proof stage, but they must be marked as author errors. Having reviewed the proofs and corrected any errors, you return them to the publisher.

The cycle of submission-review-revision and resubmission-acceptance-publication is a relatively long one. After your initial submission to a refereed journal, two months or more may pass before you receive your first feedback. After revision, another two months or more may pass (depending on whether the editor sent your paper out for a second review). The entire process (from submission to publication) can take a year or more. Chapter 3 discussed this and other publication-related issues in greater detail.

Paper Presentations

In addition to publication, you can communicate your research results through a paper presentation. Paper presentations can range from class presentations to more formal seminars to presentations at professional meetings. Two methods of presenting the results of your research are used at professional meetings. You can deliver a talk or an oral presentation before an audience, or you can present your findings in a poster session.

Oral Presentations For an *oral presentation*, you are usually given a limited amount of time to present your information. At paper sessions at professional meetings, for example, you are given 15 minutes. In that brief period of time, you must communicate to your audience the rationale behind your study, your methods, results, and conclusions. Because your time is extremely limited, it is helpful to your audience if you have a written summary of your paper (including figures and tables) to distribute at a paper session. Some conferences may require that you do this. Even if they do not, you should prepare handouts.

When preparing for an oral presentation, follow a few general rules. First, organize your talk. Do not go before a room full of professionals (or even other students) and try to "wing it." Instead, develop an outline of your talk and stick to it. Second, do not read your paper. Store what you want to say in your head. Use your notes only as a guide. Third, use appropriate visual aids whenever possible. Computerized slide shows, overhead projector transparencies, and printed handouts provide quick, easy ways to present complex methods and results. Rather than waste precious time describing methods and complex results, refer to your visual aids and restrict your discussion to the most important aspects of your research. Fourth, take some time to practice your presentation, perhaps in front of some colleagues. This will allow you to tell whether the length of your presentation is appropriate, whether you are communicating clearly, and whether you left out any elements. Feedback from your colleagues during a practice session can be valuable in identifying the strong points and weak points of your presentation.

A common mistake during oral presentations is giving an overly detailed account of the methods and procedures. It is not easy to communicate these complexities verbally. Listeners can retain only so much information and may become lost if you give too much. Try to boil down a complex method to its essential elements. To understand your study, listeners may need to know that you tested rats in an operant chamber. They may become lost in the details if you try to explain that the rats were Long–Evans females weighing between 250 and 350 grams and housed individually in 7 × 7 × 14–inch wire cages, and that the chamber was 25 × 25 × 30 centimeters (cm), constructed from aluminum sheet and fitted with a floor consisting of 0.8-cm-diameter stainless-steel bars spaced 1.2 cm apart. If your procedure is complicated, diagram it and present it on a slide or handout rather than attempting to explain it all verbally.

A final rule to follow is to avoid being pedantic and pompous. Some presenters are so consumed with self-importance that they choose to bore the audience with unnecessary details of their lives and research. Pomposity is manifested when the presenter is 13 minutes into his or her presentation and has yet to say anything about the methods and results of the study. Your audience will appreciate your presentation more and get more out of it if you focus your presentation on central issues of the research being reported.

Poster Sessions In a *poster session*, you prepare a poster that outlines the rationale behind your study and your hypotheses, method, results, and conclusions. Unlike the oral presentation, you are not limited to 15 minutes. Poster sessions may last as long as an hour or more. Many related papers are presented within each session.

The main advantage of a poster session is that you can engage in meaningful conversations with other people who are doing research in your area. During oral presentations, interaction with the audience may be limited to a few questions immediately following your presentation or a few minutes after the paper session. In a poster session, an interested person can take time to read your poster and perhaps formulate more meaningful questions and input.

Poster sessions usually require more time and effort on your part than oral presentations. Posters may be time consuming to make and difficult to transport although new technologies such as vinyl posters are cutting down on these logistical problems. Despite the difficulties, in some ways the poster format is superior to the oral presentation. It allows you greater freedom of presentation, more time for discussion, and less superficial interactions with other researchers in your area.

Conferences differ in the guidelines that you must follow to format your poster (e.g., what elements must be included, placement of sections, typeface style and size). If you are planning to present a poster, you should obtain a copy of the guidelines used at the conference at which you will be presenting. You should have your poster prepared professionally. Your university may provide services for poster preparation. If not, you can use a commercial company (e.g., an online company or Kinkos). In many cases you can have your poster professionally prepared at relatively low cost. You should not show up at your poster session with several 8.5 × 11 inch pages printed out on your home printer. Not only is this a tacky way of presenting your research, it may also be difficult for your audience to read. In addition, as was the case for the oral presentation, you should prepare a written version of your poster to distribute to interested individuals.

The Ethics of Reporting or Publishing Your Results

Your decision to report or publish your results in a public forum carries with it some important responsibilities. When you decide to publish your results, you must consider a number of ethical issues. These include avoiding plagiarism, giving proper publication credit (e.g., granting authorship status and deciding on an order of authorship), not publishing previously published results, and sharing data with qualified scientists who might want to reanalyze the data. See the APA manual (2010) for an extensive discussion of this issue.

QUESTIONS TO PONDER

1. What is typically the sequence of events involved in submitting a paper for publication?
2. What are an oral presentation and a poster session, and how do they differ?
3. What are the ethical obligations involved in reporting or publishing your results?

SUMMARY

After you have designed and conducted a study and have analyzed your data, you then prepare a report of your results. Two ways of reporting your results are the written report and the presentation. The written report establishes a permanent record of your results that is less susceptible to misunderstanding and misinterpretation than is a presentation. The presentation format gives a forum to disseminate results quickly.

Psychological research is reported using APA style (or a close variant of it), which specifies how a paper must be prepared. An APA-style paper consists of a title page, an abstract, an introduction, a method section, a results section, a discussion section, and a reference list, plus additional pages for author notes, footnotes, tables, and figures.

The title page of your paper includes the title of your paper, your name, institutional affiliation, and author notes. The abstract is a brief but concise summary of your research and is the last part of your paper that you write. The introduction of your paper introduces your topic, reviews relevant research, and states your hypotheses. It begins with a general discussion of issues and then moves to a specific discussion of your research. Its major purpose is to provide a logical justification for the study being reported.

The method section describes exactly how you conducted your study. Separate subsections provide information about participants or subjects, apparatus or materials, and procedures. The goal of the method section is to provide enough information that another researcher could replicate your study in all its essential details.

Your results are presented in the results section, which provides a detailed report of the findings (illustrated if necessary with tables and figures) and the results of any statistical analyses of the data.

The results are discussed in the discussion section. The section usually begins with a brief summary of the results about to be discussed and a restatement of any hypotheses bearing on these results. The discussion indicates whether these hypotheses were supported by the results. The findings are then related to previous knowledge in the field and conclusions are drawn.

Figures and tables are used to communicate complex information about methods or results that cannot be adequately described verbally. Graphs of relationships must be carefully drawn to avoid misleading the reader. Graphs that make small, albeit statistically significant, effects look large should not be drawn. Clear legends and captions must be included on the figure itself. Tables, like figures, can help summarize complex information. Each table consists of a number, a title, a body, and (if needed) notes describing the table. Because figures and tables are time consuming to prepare and expensive to reproduce in journals, they should only be used to illustrate complex materials or relationships.

Certain conventions are followed when you prepare an APA-style paper. References in the body of your paper are not footnoted. Instead, the author(s) of a reference is included in the text, along with the date of publication of the reference. Each reference cited in the body of your paper should be listed in expanded format in your reference section.

In a well-prepared paper, ideas must be clearly expressed, organized, and presented in a visually pleasing way. Clarity of expression requires grammatically correct sentences, properly chosen words, and economical expression. Ideas can be obscured by grammatically incorrect sentences (i.e., ambiguous sentences), poorly chosen words, and wordy sentences. Poor organization, like unclear expression, can obscure important points in your paper. Organization can be enhanced by developing and sticking to an outline.

Finally, a paper can be well written and organized yet still make a bad impression on the reader. Frequent misspellings, typographical errors, and improper word usage make your paper appear poor to your reader. Have someone proofread your paper to identify spelling and grammatical errors. Make corrections professionally by retyping rather than penciling in or crossing out information.

Plagiarism is using the words or ideas of another person without giving credit to the source. Plagiarism is considered unethical and, in a published work, may result in legal action against you. Lazy writing is writing that consists primarily of quoted (although properly cited) material. Although lazy writing is not as serious a problem as plagiarism, it is still inappropriate. To avoid both, make sure that your paper consists mainly of your own writing and ideas. Use quoted material and reference citations to support your ideas, not to present them.

After you have prepared your APA-style paper, you may want to submit it to a journal for publication. If you submit your paper to a refereed journal, it will be reviewed by outside reviewers. The editor of the journal can decide (based on the reviews) either to unconditionally accept or reject a paper or to ask you to revise and resubmit your paper. If you revise and resubmit your paper, it may be sent out for a second review. Typically, the entire cycle of initial review, second review, acceptance, and publication takes almost a year.

You also can communicate your research findings through a presentation. In an oral presentation, you stand before an audience and present your results. In a poster session, you prepare a poster and answer questions about your work. The poster session affords more time and deeper communication with others in your field than does an oral presentation.

Finally, concern over ethics does not end with the completion of your research project. Ethical practice dictates that you give proper credit to the ideas and findings of others presented in your research report, properly order the authors' names to reflect the relative importance of their contributions to the research, avoid submitting for publication findings that have already been published elsewhere, and make your data available to fellow researchers who request them.

KEY TERMS

running head
title page
author note
abstract
introduction
method section
participants subsection
subjects subsection
apparatus subsection
materials subsection
procedure subsection
results section
discussion section
reference section
biased language
plagiarism
lazy writing

APPENDIX: STATISTICAL TABLES

TABLE 1A 1,000 Six-Digit Random Numbers

TABLE 1B Random Orderings of the Numbers 1–30

TABLE 2 Critical Values of *t*

TABLE 3A Critical Values of *F* ($p < .05$)

TABLE 3B Critical Values of *F* ($p < .01$)

TABLE 4A Critical Values of the Mann–Whitney *U* Test ($p < .05$, two-tailed test)

TABLE 4B Critical Values of the Mann–Whitney *U* Test ($p < .01$, two-tailed test)

TABLE 5 Areas Under the Normal Curve

TABLE 6 Critical Values of Chi-Square

TABLE 7 Conversion of *r* to *z*

TABLE 1A **1,000 Six-Digit Random Numbers (first 500)**

1	2	3	4	5	6	7	8	9	10
192805	905642	577821	582703	418793	921234	423676	926116	359852	611072
696843	580817	915407	920290	587586	090028	592469	094911	428558	679779
459327	255992	252995	257878	756380	258822	761262	263704	528645	779865
706608	962545	590582	595465	925173	427615	930056	432498	597352	848572
603541	637720	928169	933052	093968	596408	098850	601291	697438	948659
469093	312895	265757	270640	262761	765202	267644	770085	766145	017367
283598	988069	634724	639607	431554	933995	436437	938878	866232	117453
716374	663244	972311	977194	600348	102790	605231	107673	934939	186160
180530	338419	309899	314782	769141	271583	774024	276466	035026	286247
613306	044973	647486	652369	937935	440377	942818	445260	103733	354954
046083	720147	985073	989956	106729	609170	116612	614053	203820	455040
478858	395322	322661	327543	275523	777964	280406	782847	272527	523747
561285	070497	660248	665130	475696	978137	480579	983020	372613	623834
293364	745671	997834	002718	644490	146932	649372	151814	441320	692541
025442	420846	335422	340305	813283	315725	818166	320608	541407	792627
726139	127400	673009	677892	982077	484518	986959	489401	610114	861334
458218	802574	010597	015480	150871	653312	155754	658195	710200	961421
190296	477749	348184	353067	319664	822105	324547	826988	778907	030218
890993	152924	717151	722034	488458	990899	493341	995782	878994	130215
623072	828098	054739	059622	657251	159693	662134	164576	947700	198922
323770	503273	392326	397209	826045	328487	830928	333370	047788	299009
055848	209827	729913	734796	994838	497280	999721	502163	116495	367715
787926	885001	067501	072383	163633	666074	168516	670956	216582	467802
488624	560176	405088	409970	332426	834867	337309	839750	285288	536509
220702	235351	742675	747557	501220	003662	506103	008544	385375	636596
571051	910525	080262	085145	670013	172455	674896	177338	485462	736682
952780	585700	417849	422732	838807	341249	843689	346131	554169	805389
303129	260874	755436	760319	007601	510042	012484	514925	654255	905476
653478	967428	093024	097907	176395	678835	181277	683718	722962	974182
035208	642603	430611	435494	345188	847629	350071	852512	823049	074270
385556	317778	768198	773081	513981	016423	518864	021306	891755	142977
735905	992952	137166	142049	682775	185217	687658	190100	991842	243064
117635	668127	474753	479636	851568	354010	856451	358893	060550	311770
467983	343301	812340	817223	020363	522804	025246	527687	160637	411857
818332	049856	149928	154810	189156	691597	194039	696480	229343	480564
200062	725030	487515	492397	357950	860391	362833	865273	329430	580651
550410	400205	825102	829984	562743	029185	531626	034068	398137	649357
900759	075380	162689	167572	726917	229359	731799	234241	498224	749444
251109	750554	500276	505159	895710	398152	900593	403035	566930	818151
632837	425728	837863	842746	064505	566945	069387	571828	667017	918237
983186	132283	175451	180334	233298	735739	238181	740622	735724	986944
333536	807457	513038	517921	402091	904532	406974	909415	835810	087032
715264	482632	850625	855508	570885	073327	575768	078210	904517	155739
065614	157807	219593	224476	739678	242120	744561	247003	004605	255825
415963	832980	557180	562063	908472	410914	913355	415797	073312	324532
797691	508155	894767	899650	077266	579707	082149	584590	173398	424619
148041	214710	232355	237237	246060	748501	250943	753383	242105	493326
498390	889884	569942	574824	414853	917294	419736	922177	342192	593412
880118	565059	907529	912411	583647	086089	588529	090971	410899	662119
230468	240234	245116	249999	752440	254882	757323	259765	510985	762206

TABLE 1A **1,000 Six-Digit Random Numbers (second 500)**

1	2	3	4	5	6	7	8	9	10
869354	895876	568055	572938	409027	911468	413910	916351	350086	601306
687077	571051	905642	910525	577821	080262	582703	085145	418793	670013
449562	246226	243230	248113	746614	249056	751497	253939	518879	770100
696843	952780	580817	585700	915407	417849	920290	422732	587586	838807
593775	627955	918404	923286	084202	586643	089085	591526	687673	938893
459327	303129	255992	260874	252995	755436	457878	760319	756380	007601
273833	978303	624958	629841	421789	924230	426672	929113	856466	107688
706608	653478	962545	967428	590582	093024	595465	097907	925173	176395
170765	328653	300133	305016	759376	261818	764259	266701	025261	276481
603541	035208	637720	642603	928169	430611	933052	435494	093968	345188
036317	710382	975307	980190	096964	599405	101847	604287	194054	445275
469093	385556	312895	317778	265757	768198	270640	773081	262761	513981
551520	060731	650482	655365	465931	968371	470813	973254	362848	614068
283598	735905	988069	992952	634724	137166	639607	142049	431554	682775
015676	411080	325657	330540	803517	305959	808400	310842	531641	782862
716374	117635	663244	668127	972311	474753	977194	470636	600348	851568
448452	792809	000832	005714	141105	643546	145988	648429	700435	951655
180530	467983	338419	343301	309899	812340	314782	817223	769141	020363
881228	143158	707385	712268	478692	981133	483575	986016	869228	120450
613306	818332	044973	049856	647486	149928	652369	154810	937935	189156
314005	493507	382560	387443	816279	318721	821162	323604	038023	289243
046083	200062	720147	725030	985073	487515	989956	492397	106729	357950
778160	875236	057735	062618	153867	656308	158750	611191	206816	458036
478858	550410	395322	400205	322661	825102	327543	829984	275523	526743
210937	225585	732909	737792	491454	993895	496337	998778	375609	626830
561285	900759	070497	075380	660248	162689	665130	167572	475696	726917
943014	575934	408084	412967	829041	331483	833924	336366	544403	795623
293364	251109	745671	750554	997834	500276	002718	505159	644490	895710
643712	957663	083259	088141	166629	669070	171512	673953	713196	964417
025442	632837	420846	425728	335422	837863	340305	842746	813283	064505
375791	308012	758432	763315	504216	006658	509099	011541	881990	133211
726139	983186	127400	132283	673009	175451	677892	180334	982077	233298
107869	658361	464987	469870	841803	344245	846686	349127	050784	302005
458218	333536	802574	807457	010597	513038	015480	517921	150871	402091
808566	040090	140162	145045	179391	681832	184274	686714	219578	470798
190296	715264	477749	482632	348184	850625	353067	855508	319664	570885
540645	390439	815336	820219	516978	019420	521860	024302	388371	639592
890993	065614	152924	157807	717151	219593	722034	224476	488458	739678
241343	740788	490511	495394	855944	388386	890827	393269	557165	808385
623072	415963	828098	832980	054739	557180	059622	562063	657251	908472
973420	122517	165686	170568	223532	725973	228415	730856	725958	977179
323770	797691	503273	508155	392326	894767	397209	899650	826045	077266
705499	472866	840559	845742	561119	063561	566002	068444	894752	145973
055848	148041	209827	214710	729913	232355	734796	237237	994838	246060
406197	823215	547414	552297	898706	401148	903589	406031	063546	314767
787926	498390	885001	889884	067501	569942	072383	574824	163633	414853
138275	204944	222589	227472	236294	738735	241177	743618	232340	483560
488624	880118	560176	565059	405088	907529	409970	912411	332426	583647
870353	555293	897763	902646	573881	076323	578764	081206	401133	652353
220702	230468	235351	240234	742675	245116	747557	249999	501220	752440

NOTE: This table was generated with a computer program written in BASIC.

TABLE 1B Random Orderings of the Numbers 1–30

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
25	15	25	30	20	14	25	3	21	2	20	27	19	26	17	30	22	4	6	10
14	26	17	7	2	9	14	13	6	20	29	16	13	19	9	17	14	16	22	20
17	11	2	17	21	13	3	24	23	21	12	24	24	2	3	25	6	17	25	29
5	4	18	18	23	20	16	20	26	14	1	7	26	21	22	9	26	13	27	22
3	16	12	19	16	22	10	7	13	7	21	20	29	11	24	20	27	8	12	15
19	20	5	11	11	26	6	21	16	22	23	21	20	7	16	22	20	22	17	25
22	6	23	21	24	17	19	22	29	25	10	4	6	22	13	5	19	24	29	26
10	22	24	24	18	1	21	23	19	23	24	23	22	25	26	23	30	26	20	19
23	23	16	14	17	19	11	16	2	15	27	9	3	16	27	15	21	28	1	28
24	13	28	25	30	5	26	26	20	26	16	25	23	27	21	28	2	3	23	21
16	28	27	28	19	21	15	28	3	28	30	26	11	28	2	27	23	21	3	3
27	29	1	2	5	10	2	30	22	29	22	18	25	30	4	19	10	5	24	4
28	3	21	5	9	23	4	1	10	30	4	28	16	4	23	29	25	23	10	23
29	21	4	22	22	25	20	6	24	4	6	29	27	24	5	2	15	10	26	11
30	5	7	9	14	16	9	25	15	10	25	30	28	10	25	7	28	27	16	27
4	10	26	26	25	27	24	12	27	27	14	1	2	15	15	26	1	14	30	16
8	27	13	15	1	30	13	27	1	19	28	5	5	29	28	12	3	29	2	1
13	14	29	1	3	3	27	17	4	1	18	13	9	18	30	16	9	30	4	7
2	30	30	4	7	6	30	2	8	3	2	17	15	5	19	4	11	20	8	8
21	19	20	8	13	12	17	4	11	8	7	2	1	8	7	6	29	6	28	14
6	9	6	12	15	15	7	9	28	12	13	6	4	14	11	10	18	11	13	2
11	12	9	29	4	4	22	15	17	18	17	11	8	1	1	13	4	1	21	5
1	2	14	16	8	7	23	19	5	5	5	14	12	20	20	3	8	19	5	9
20	17	3	6	12	24	1	5	9	9	8	3	30	9	6	21	12	7	9	12
7	7	22	23	26	2	18	10	12	11	11	22	18	12	10	11	16	12	14	30
12	24	10	13	28	18	8	14	30	16	15	10	21	17	14	14	5	15	19	18
15	1	15	3	29	8	12	18	18	6	19	15	10	6	18	18	24	18	7	6
18	18	19	20	6	11	28	8	7	24	9	19	14	23	8	8	13	9	11	24
9	8	8	10	10	28	29	11	25	13	26	8	17	13	12	24	17	25	15	13
26	25	11	27	27	29	5	29	14	17	3	12	7	3	29	1	7	2	18	17

TABLE 1B **Random Orderings of the Numbers 1–30** ***continued***

21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
19	19	29	5	19	15	15	21	18	8	11	6	6	11	30	25	16	20	19	11
26	12	21	17	30	11	8	9	7	14	3	17	18	24	16	13	23	26	15	12
11	21	5	24	11	24	20	22	16	23	15	19	13	21	6	18	28	8	26	27
8	24	24	19	4	28	25	25	24	28	20	10	11	15	20	1	20	3	30	22
21	27	17	8	24	17	27	19	26	17	19	15	21	2	22	4	13	24	22	17
17	30	10	22	22	22	28	16	20	20	6	20	25	16	12	21	2	16	23	1
13	22	25	16	9	4	22	27	19	3	26	23	27	5	23	22	21	15	5	20
28	5	27	12	27	21	1	29	6	18	24	25	28	19	26	8	5	27	24	3
29	23	19	25	25	6	23	20	22	7	25	27	1	20	18	23	22	30	9	21
18	15	30	28	17	23	4	3	13	21	18	3	22	8	27	24	10	29	25	7
2	25	23	30	28	12	24	5	23	11	28	22	3	23	28	15	25	21	18	23
6	18	2	1	1	25	12	23	17	24	29	7	24	13	2	27	27	23	27	24
22	28	9	3	23	27	26	12	25	27	30	24	8	26	4	28	19	7	28	15
10	29	26	23	2	16	17	26	2	15	5	14	26	29	25	30	29	11	2	29
27	2	14	7	6	29	29	17	3	29	9	26	17	1	9	2	30	28	4	30
15	7	28	26	26	30	30	28	21	30	27	16	30	4	13	6	4	19	8	2
30	11	18	15	15	5	2	1	10	5	16	4	2	7	29	11	7	1	13	4
3	16	3	29	29	7	6	7	14	10	1	9	4	25	19	29	12	4	1	9
5	1	7	18	21	14	11	11	1	13	22	13	9	12	5	17	15	9	3	28
9	4	13	6	3	1	16	15	4	1	8	1	29	30	10	3	3	14	6	14
12	9	1	9	7	19	3	2	9	4	12	5	16	17	15	5	6	2	11	19
16	13	20	13	12	9	5	6	11	22	2	21	20	6	3	10	11	5	29	6
7	17	8	2	18	13	9	10	15	12	21	11	7	9	21	14	14	25	16	10
23	6	11	20	5	3	13	13	5	2	10	2	10	27	11	20	17	12	21	13
25	8	15	11	8	18	19	30	8	19	13	18	14	28	14	7	8	18	7	18
14	26	4	14	13	8	7	18	12	9	17	8	5	3	17	12	24	6	12	8
4	3	22	4	16	26	10	8	27	25	7	12	23	22	8	16	1	10	17	25
20	20	12	21	20	2	14	24	28	26	23	28	12	10	24	19	18	13	20	26
24	10	16	10	10	20	18	14	29	16	14	29	15	14	1	9	9	17	10	16
1	14	6	27	14	10	21	4	30	6	4	30	19	18	7	26	26	22	14	5

continues

TABLE 1B **Random Orderings of the Numbers 1–30** ***continued***

41	42	43	44	45	46	47	48	49	50
20	15	25	25	12	7	16	17	7	1
16	19	21	20	30	27	8	30	15	17
9	3	1	3	17	15	25	20	28	18
26	11	22	21	7	23	28	23	25	14
23	21	13	13	19	3	23	9	16	3
24	23	9	7	22	18	20	26	3	20
18	18	23	22	11	21	1	24	19	10
28	17	24	23	24	5	21	13	22	11
29	24	18	16	25	22	5	29	5	21
22	26	26	26	16	12	24	28	24	24
4	30	28	28	28	24	13	19	11	27
7	6	29	30	18	25	26	21	26	28
25	22	30	1	1	17	27	3	27	2
14	10	3	6	5	28	18	6	18	4
27	14	5	24	8	29	29	11	29	22
19	27	27	12	26	30	30	27	30	7
2	1	14	27	15	4	2	18	4	26
5	4	20	17	29	11	6	1	10	13
11	7	2	2	3	14	10	8	14	29
15	12	6	4	21	1	17	10	1	30
3	28	10	9	9	19	3	16	21	19
6	16	17	15	13	8	7	2	8	9
10	5	4	19	2	13	11	7	12	12
13	8	7	5	6	16	14	25	2	16
30	25	11	10	23	6	4	14	20	5
17	2	15	14	14	9	22	5	9	8
21	20	19	18	4	26	12	22	13	25
12	9	8	8	20	2	15	12	17	15
1	13	12	11	10	20	19	15	6	6
8	29	16	29	27	10	9	4	23	23

NOTE: These random orders were derived with a computer program written in BASIC.

TABLE 2 Critical Values of *t*

Degrees of Freedom	ALPHA LEVEL (TWO-TAILED TEST)									
	.8	.5	.2	.1	.05	.02	.01	.005	.002	.001
1	0.325	1.00	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.598
3	.277	.765	1.638	2.353	3.182	4.541	5.841	7.453	10.214	12.924
4	.271	.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	.265	.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	.263	.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	.262	.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	.261	.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	.260	.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	.259	.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	.259	.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	.258	.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	.258	.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	.257	.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	.257	.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	.257	.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	.257	.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	.256	.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	.256	.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	.256	.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	.256	.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	.256	.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	.256	.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	.256	.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	.255	.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	.254	.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	.254	.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	.253	.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

NOTE: To obtain one-tailed alpha levels, simply divide the two-tailed alpha by 2 (for example, $p < .05$ for a one-tailed test is 1/2 = .05).

SOURCE: Adapted from Table 12, *Biometrika: Tables for Statisticians* (Vol. 1, 3rd ed.), 1966, by E. S. Pearson & H. O. Hartley; reprinted with permission.

TABLE 3A Critical Values of *F* (*p* < .05)

Degrees of Freedom in the Denominator	DEGREES OF FREEDOM IN THE NUMERATOR								
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88

SOURCE: Adapted from Table 18, *Biometrika: Tables for Statisticians* (Vol. 1, 3rd ed.), 1966, by E. S. Pearson & H. O. Hartley; reprinted with permission.

TABLE 3A Critical Values of F ($p < .05$) *continued*

	DEGREES OF FREEDOM IN THE NUMERATOR									
Degrees of Freedom in the Denominator	10	12	15	20	24	30	40	60	120	∞
1	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

TABLE 3B **Critical Values of F (p < .01)**

	DEGREES OF FREEDOM IN THE NUMERATOR								
Degrees of Freedom in the Denominator	1	2	3	4	5	6	7	8	9
1	4,052	4,999.5	5,403	5,625	5,764	5,859	5,928	5,981	6,022
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22
26	7.72	5.53	5.64	4.14	3.82	3.59	3.42	3.29	3.18
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41

SOURCE: Adapted from Table 18, *Biometrika: Tables for Statisticians* (Vol. 1, 3rd ed.), 1966, by E. S. Pearson & H. O. Hartley; reprinted with permission.

TABLE 3B Critical Values of F ($p < .01$)

Degrees of Freedom in the Denominator	DEGREES OF FREEDOM IN THE NUMERATOR									
	10	*12*	*15*	*20*	*24*	*30*	*40*	*60*	*120*	∞
1	6,056	6,106	6,157	6,209	6,235	6,261	6,287	6,313	6,339	6,366
2	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

TABLE 4A Critical Values of the Mann–Whitney *U* Test

	$p < .05$ (TWO-TAILED TEST)																			
	n																			
m	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***13***	***14***	***15***	***16***	***17***	***18***	***19***	***20***
1	—																			
2	—	—																		
3	—	—	—																	
4	—	—	—	0																
5	—	—	0	1	2															
6	—	—	1	2	3	5														
7	—	—	1	3	5	6	8													
8	—	0	2	4	6	8	10	13												
9	—	0	2	4	7	10	12	15	17											
10	—	0	3	5	8	11	14	17	20	23										
11	—	0	3	6	9	13	16	19	23	26	30									
12	—	1	4	7	11	14	18	22	26	29	33	37								
13	—	1	4	8	12	16	20	24	28	33	37	41	45							
14	—	1	5	9	13	17	22	26	31	36	40	45	50	55						
15	—	1	5	10	14	19	24	29	34	39	44	49	54	59	64					
16	—	1	6	11	15	21	26	31	37	42	47	53	59	64	70	75				
17	—	2	6	11	17	22	28	34	39	45	51	57	63	69	75	81	87			
18	—	2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99		
19	—	2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	
20	—	2	8	14	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127
21	—	3	8	15	22	29	36	43	50	58	65	73	80	88	96	103	111	119	126	134
22	—	3	9	16	23	30	38	45	53	61	69	77	85	93	101	109	117	125	133	141
23	—	3	9	17	24	32	40	48	56	64	73	81	89	98	106	115	123	132	140	149
24	—	3	10	17	25	33	42	50	59	67	76	85	94	102	111	120	129	138	147	156
25	—	3	10	18	27	35	44	53	62	71	80	89	98	107	117	126	135	145	154	163
26	—	4	11	19	28	37	46	55	64	74	83	93	102	112	122	132	141	151	161	171
27	—	4	11	20	29	38	48	57	67	77	87	97	107	117	127	137	147	158	168	178
28	—	4	12	21	30	40	50	60	70	80	90	101	111	122	132	143	154	164	175	186
29	—	4	13	22	32	42	52	62	73	83	94	105	116	127	138	149	160	171	182	193
30	—	5	13	23	33	43	54	65	76	87	98	109	120	131	143	154	166	177	189	200
31	—	5	14	24	34	45	56	67	78	90	101	113	125	136	148	160	172	184	196	208
32	—	5	14	24	35	46	58	69	81	93	105	117	129	141	153	166	178	190	203	215
33	—	5	15	25	37	48	60	72	84	96	108	121	133	146	159	171	184	197	210	222
34	—	5	15	26	38	50	62	74	87	99	112	125	138	151	164	177	190	203	217	230
35	—	6	16	27	39	51	64	77	89	103	116	129	142	156	169	183	196	210	224	237
36	—	6	16	28	40	53	66	79	92	106	119	133	147	161	174	188	202	216	231	245
37	—	6	17	29	41	55	68	81	95	109	123	137	151	165	180	194	209	223	238	252
38	—	6	17	30	43	56	70	84	98	112	127	141	156	170	185	200	215	230	245	259
39	0	7	18	31	44	58	72	86	101	115	130	145	160	175	190	206	221	236	252	267
40	0	7	18	31	45	59	74	89	103	119	134	149	165	180	196	211	227	243	258	274

SOURCE: Reprinted from R. C. Milton (1964), An extended table of critical values for the Mann–Whitney (Wilcoxon) two-sample statistic, *Journal of the American Statistical Association, 59,* 925–934.

TABLE 4B **Critical Values of the Mann–Whitney *U* Test**

	$p < .01$ (TWO-TAILED TEST)																			
	n																			
m	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	—																			
2	—	—																		
3	—	—	—																	
4	—	—	—	—																
5	—	—	—	—	0															
6	—	—	—	0	1	2														
7	—	—	—	0	1	3	4													
8	—	—	—	1	2	4	6	7												
9	—	—	0	1	3	5	7	9	11											
10	—	—	0	2	4	6	9	11	13	16										
11	—	—	0	2	5	7	10	13	16	18	21									
12	—	—	1	3	6	9	12	15	18	21	24	27								
13	—	—	1	3	7	10	13	17	20	24	27	31	34							
14	—	—	1	4	7	11	15	18	22	26	30	34	38	42						
15	—	—	2	5	8	12	16	20	24	29	33	37	42	46	51					
16	—	—	2	5	9	13	18	22	27	31	36	41	45	50	55	60				
17	—	—	2	6	10	15	19	24	29	34	39	44	49	54	60	65	70			
18	—	—	2	6	11	16	21	26	31	37	42	47	53	58	64	70	75	81		
19	—	0	3	7	12	17	22	28	33	39	45	51	57	63	69	74	81	87	93	
20	—	0	3	8	13	18	24	30	36	42	48	54	60	67	73	79	86	92	99	105
21	—	0	3	8	14	19	25	32	38	44	51	58	64	71	78	84	91	98	105	112
22	—	0	4	9	14	21	27	34	40	47	54	61	68	75	82	89	96	104	111	118
23	—	0	4	9	15	22	29	35	43	50	57	64	72	79	87	94	102	109	117	125
24	—	0	4	10	16	23	30	37	45	52	60	68	75	83	91	99	107	115	123	131
25	—	0	5	10	17	24	32	39	47	55	63	71	79	87	96	104	112	121	129	138
26	—	0	5	11	18	25	33	41	49	58	66	74	83	92	100	109	118	127	135	144
27	—	1	5	12	19	27	35	43	52	60	69	78	87	96	105	114	123	132	142	151
28	—	1	5	12	20	28	36	45	54	63	72	81	91	100	109	119	128	138	148	157
29	—	1	6	13	21	29	38	47	56	66	75	85	94	104	114	124	134	144	154	164
30	—	1	6	13	22	30	40	49	58	68	78	88	98	108	119	129	139	150	160	170
31	—	1	6	14	22	32	41	51	61	71	81	92	102	113	123	134	145	155	166	177
32	—	1	7	14	23	33	43	53	63	74	84	95	106	117	128	139	150	161	172	184
33	—	1	7	15	24	34	44	55	65	76	87	98	110	121	132	144	155	167	179	190
34	—	1	7	16	25	35	46	57	68	79	90	102	113	125	137	149	161	173	185	197
35	—	1	8	16	26	37	47	59	70	82	93	105	117	129	142	154	166	179	191	203
36	—	1	8	17	27	38	49	60	72	84	96	109	121	134	146	159	172	184	197	210
37	—	1	8	17	28	39	51	62	75	87	99	112	125	138	151	164	177	190	203	217
38	—	1	9	18	29	40	52	64	77	90	102	116	129	142	155	169	182	196	210	223
39	—	2	9	19	30	41	54	66	79	92	106	119	133	146	160	174	188	202	216	230
40	—	2	9	19	31	43	55	68	81	95	109	122	136	150	165	179	193	208	222	237

SOURCE: Reprinted from R. C. Milton (1964), An extended table of critical values for the Mann–Whitney (Wilcoxon) two-sample statistic, *Journal of the American Statistical Association, 59,* 925–934.

TABLE 5 Areas Under the Normal Curve

Ones and Tenths Value of z	HUNDREDTHS VALUE OF z									
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2624	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.49903									
3.2	.49931									
3.3	.49952									
3.4	.49966									
3.5	.49977									
3.6	.49984									
3.7	.49989									
3.8	.49993									
3.9	.49995									
4.0	.50000									

SOURCE: Reprinted with permission from *Computational Handbook of Statistics,* by J. L. Bruning & B. L. Kintz. Copyright © 1987, 1977, 1968 by Scott, Foresman & Company.

TABLE 6 Critical Values of Chi-Square

	p VALUE						
Degrees of Freedom	.25	.10	.05	.025	.01	.005	.001
1	1.32330	2.70554	3.84146	5.02389	6.63490	7.87944	10.828
2	2.77259	4.60517	5.99146	7.37776	9.21034	10.5966	13.816
3	4.10834	6.25139	7.81473	9.34840	11.3449	12.8382	16.266
4	5.38527	7.77944	9.48773	11.1433	13.2767	14.8603	18.467
5	6.62568	9.23636	11.0705	12.8325	15.0863	16.7496	20.515
6	7.84080	10.6446	12.5916	14.4494	16.8119	18.5476	22.458
7	9.03715	12.0170	14.0671	16.0128	18.4753	20.2777	24.322
8	10.2189	13.3616	15.5073	17.5345	20.0902	21.9550	26.125
9	11.3888	14.6837	16.9190	19.0228	21.6660	23.5894	27.877
10	12.5489	15.9872	18.3070	20.4832	23.2093	25.1882	29.588
11	13.7007	17.2750	19.6751	21.9200	24.7250	26.7568	31.264
12	14.8454	18.5493	21.0261	23.3367	26.2170	28.2995	32.909
13	15.9839	19.8119	22.3620	24.7356	27.6882	29.8195	34.528
14	17.1169	21.0641	23.6848	26.1189	29.1412	31.3194	36.123
15	18.2451	22.3071	24.9958	27.4884	30.5779	32.8013	37.697
16	19.3689	23.5418	26.2962	28.8454	31.9999	34.2672	39.252
17	20.4887	24.7690	27.5871	30.1910	33.4087	35.7185	40.790
18	21.6049	25.9894	28.8693	31.5264	34.8053	37.1565	42.312
19	22.7178	27.2036	30.1435	32.8523	36.1909	38.5823	43.820
20	23.8277	28.4120	31.4104	34.1696	37.5662	39.9968	45.315
21	24.9348	29.6151	32.6706	35.4789	38.9322	41.4011	46.797
22	26.0393	30.8133	33.9244	36.7807	40.2894	42.7957	48.268
23	27.1413	32.0069	35.1725	38.0756	41.6384	44.1813	49.728
24	28.2412	33.1962	36.4150	39.3641	42.9798	45.5585	51.179
25	29.3389	34.3816	37.6525	40.6465	44.3141	46.9279	52.618
26	30.4346	35.5632	38.8851	41.9232	45.6417	48.2899	54.052
27	31.5284	36.7412	40.1133	43.1945	46.9629	49.6449	55.476
28	32.6205	37.9159	41.3371	44.4608	48.2782	50.9934	56.892
29	33.7109	39.0875	42.5570	45.7223	49.5879	52.3356	58.301
30	34.7997	40.2560	43.7730	46.9792	50.8922	53.6720	59.703
40	45.6160	51.8051	55.7585	59.3417	63.6907	66.7660	73.402
50	56.3336	63.1671	67.5048	71.4202	76.1539	79.4900	86.661
60	66.9815	74.3970	79.0819	83.2977	88.3794	91.9517	99.607
70	77.5767	85.5270	90.5312	95.0232	100.425	104.215	112.317
80	88.1303	96.5782	101.879	106.629	112.329	116.321	124.839
90	98.6499	107.565	113.145	118.136	124.116	218.299	137.208
100	109.141	118.498	124.342	129.561	135.807	140.169	149.449

SOURCE: Adapted from Table 8, *Biometrika: Tables for Statisticians* (Vol. 1, 3rd ed.), 1966, by E. S. Pearson & H. O. Hartley; reprinted with permission.

TABLE 7 Conversion of *r* to *z*

r	*z*	*r*	*z*	*r*	*z*	*r*	*z*	*r*	*z*
.000	.000	.200	.203	.400	.424	.600	.693	.800	1.099
.005	.005	.205	.208	.405	.430	.605	.701	.805	1.113
.010	.010	.210	.213	.410	.436	.610	.709	.810	1.127
.015	.015	.215	.218	.415	.442	.615	.717	.815	1.142
.020	.020	.220	.224	.420	.448	.620	.725	.820	1.157
.025	.025	.225	.229	.425	.454	.625	.733	.825	1.172
.030	.030	.230	.234	.430	.460	.630	.741	.830	1.188
.035	.035	.235	.239	.435	.466	.635	.750	.835	1.204
.040	.040	.240	.245	.440	.472	.640	.758	.840	1.221
.045	.045	.245	.250	.445	.478	.645	.767	.845	1.238
.050	.050	.250	.255	.450	.485	.650	.775	.850	1.256
.055	.055	.255	.261	.455	.491	.655	.784	.855	1.274
.060	.060	.260	.266	.460	.497	.660	.793	.860	1.293
.065	.065	.265	.271	.465	.504	.665	.802	.865	1.313
.070	.070	.270	.277	.470	.510	.670	.811	.870	1.333
.075	.075	.275	.282	.475	.517	.675	.820	.875	1.354
.080	.080	.280	.288	.480	.523	.680	.829	.880	1.376
.085	.085	.285	.293	.485	.530	.685	.838	.885	1.398
.090	.090	.290	.299	.490	.536	.690	.848	.890	1.422
.095	.095	.295	.304	.495	.543	.695	.858	.895	1.447
.100	.100	.300	.310	.500	.549	.700	.867	.900	1.472
.105	.105	.305	.315	.505	.556	.705	.877	.905	1.499
.110	.110	.310	.321	.510	.563	.710	.887	.910	1.528
.115	.116	.315	.326	.515	.570	.715	.897	.915	1.557
.120	.121	.320	.332	.520	.576	.720	.908	.920	1.589
.125	.126	.325	.337	.525	.583	.725	.918	.925	1.623
.130	.131	.330	.343	.530	.590	.730	.929	.930	1.658
.135	.136	.335	.348	.535	.597	.735	.940	.935	1.697
.140	.141	.340	.354	.540	.604	.740	.950	.940	1.738
.145	.146	.345	.360	.545	.611	.745	.962	.945	1.783
.150	.151	.350	.365	.550	.618	.750	.973	.950	1.832
.155	.156	.355	.371	.555	.626	.755	.984	.955	1.886
.160	.161	.360	.377	.560	.633	.760	.996	.960	1.946
.165	.167	.365	.383	.565	.640	.765	1.008	.965	2.014
.170	.172	.370	.388	.570	.648	.770	1.020	.970	2.092
.175	.177	.375	.394	.575	.655	.775	1.033	.975	2.185
.180	.182	.380	.400	.580	.662	.780	1.045	.980	2.298
.185	.187	.385	.406	.585	.670	.785	1.058	.985	2.443
.190	.192	.390	.412	.590	.678	.790	1.071	.990	2.647
.195	.198	.395	.418	.595	.685	.795	1.085	.995	2.994

SOURCE: Reprinted with permission from A. L. Edwards (1985), *Experimental Design in Psychological Research* (5th ed.).

GLOSSARY

ABAB design In a single-subject baseline design, the baseline (A) and intervention (B) phases are each repeated to provide an immediate intrasubject replication.

abstract A concise summary of an APA-style manuscript that includes a brief description of the rationale for the study, methods, results, and conclusions.

accuracy Agreement of a measurement with a known standard.

alpha level (α) The probability of obtaining a difference at least as large as the one actually obtained, given that the difference occurred purely as a result of chance factors. By convention, the maximum acceptable alpha level is .05 (5 chances in 100 or 1 chance in 20).

analogical theory A theory that explains a relationship through analogy to a well-understood model.

analysis of covariance (ANCOVA) Variant of the analysis of variance used to analyze data from experiments that include a correlational variable (covariate).

analysis of variance (ANOVA) An inferential statistic used to evaluate data from experiments with more than two levels of an independent variable or data from multifactor experiments. Versions are available for between-subjects and within-subjects designs.

apparatus subsection Subsection of the method section of an APA-style manuscript in which any equipment, materials, and measures are described in detail. Sometimes called the *materials* subsection.

applied research Research carried out to investigate a real-world problem.

archival research A nonexperimental research strategy in which you make use of existing records as your basic source for data.

author note An element of the title page providing author, departmental affiliation, acknowledgments, disclaimers or conflicts of interest, and author contact information.

bar graph A graph on which data from groups of subjects are represented by bars of differing heights tied to the value of the dependent variable for the group.

baseline design A single-subject experimental design in which subjects are observed under each of several treatment conditions. Observations made during baseline periods (no treatment) are compared with observations made during intervention periods (treatment introduced).

baseline phase Phase of a single-subject, baseline design in which you establish the level of performance on the dependent measure before introducing the treatment.

basic research Research carried out primarily to test a theory or empirical issues.

behavioral baseline Level of behavior under the baseline and intervention phases of a single-subject, baseline design. It is used to determine the amount of uncontrolled variability in the data.

behavioral categories The general and specific classes of behavior to be observed in an observational study.

behavioral measure A measure of a subject's activity in a situation, for example, the number of times a rat presses a lever (frequency of responding).

belief-based explanation An explanation for behavior that is accepted without evidence because it comes from a trusted source or fits within a larger framework of belief.

Belmont Report A report issued in 1979 presenting three basic principles of ethical treatment of human participants that underlie all medical and behavioral research (respect for persons, beneficence, and justice).

beneficence An ethical principle in the Belmont Report stating that researchers will do no harm to participants and will strive to maximize benefits to the participant while minimizing harm.

beta weight (β) Standardized regression weight used to interpret the results of a linear regression analysis. A beta weight can be interpreted as a partial correlation coefficient.

between-subjects design An experimental design in which different groups of subjects are exposed to the various levels of the independent variable.

biased sample A sample that is not representative of the population it is supposed to represent.

bivariate linear regression A statistical technique for fitting a straight line to a set of data points representing the paired values of two variables.

boxplot A graphical display of the values of the five-number summary of a distribution.

canonical correlation Multivariate statistical techniques used to correlate two sets of variables.

carryover effect A problem associated with within-subjects designs in which exposure to one level of the independent variable alters the behavior observed under subsequent levels.

case history A nonexperimental research technique in which an individual case is studied intensively to uncover its history (e.g., a patient in therapy).

causal relationship A relationship in which changes in the value of one variable cause changes in the value of another.

chi-square (χ^2) Nonparametric inferential statistic used to evaluate the relationship between variables measured on a nominal scale.

circular explanation (or tautology) An explanation of behavior that refers to factors whose only proof of existence is the behavior they are being called on to explain.

cluster sampling A sampling technique in which naturally occurring groups (such as students in an elementary school class) are randomly selected for inclusion in a sample.

coefficient of nondetermination Statistic indicating the proportion of variance in one variable not accounted for by variation in a second variable.

Cohen's Kappa A popular statistic used to assess interrater reliability. It compares the observed proportion of agreement to the proportion of agreement that would be expected if agreement occurred purely by chance.

cohort-sequential design A developmental design including cross-sectional and longitudinal components.

commonsense explanations Loose explanations for behavior that are based on what we believe to be true about the world.

concurrent validity The validity of a test established by showing that its results can be used to infer an individual's value on some other, accepted test administered at the same time.

confirmational strategy A strategy for testing a theory that involves finding evidence that confirms the predictions made by the theory.

confirmation bias The human tendency to seek out information that confirms what is already believed.

confounding Two variables that vary together in such a way that the effects of one cannot be separated from the effects of the other.

construct validity Validity that applies when a test is designed to measure a "construct" or variable "constructed" to describe or explain behavior on the basis of theory (e.g., intelligence). A test has construct validity if the measured values of the construct predict behavior as expected from the theory (e.g., those with higher intelligence scores achieve higher grades in school).

content analysis A nonexperimental research technique that is used to analyze a written or spoken record for the occurrence of specific categories of events.

content validity Validity of a test established by judging how adequately the test samples behavior representative of the universe of behaviors the test was designed to sample.

control group A group of subjects in an experiment that does not receive the experimental treatment. The data from the control group are used as a baseline against which data from the experimental group are compared.

correlational relationship A relationship in which the value of one variable changes systematically with the value of a second variable.

correlational research Research in which no independent variables are manipulated. Instead, two or more dependent variables are measured to identify possible correlational relationships.

correlation matrix A matrix giving the set of all possible bivariate correlations among three or more variables.

counterbalancing A technique used to combat carryover effects in within-subjects designs. Counterbalancing involves assigning the various treatments of an experiment in a different order for different subjects.

covariate A correlational variable (usually a characteristic of the subject) included in an experiment to help reduce the error variance in statistical tests.

criterion-related validity The ability of a measure to produce results similar to those provided by other, established measures of the same variable.

critical region Portion of the sampling distribution of a statistic within which observed values of the statistic are considered to be statistically significant. Usually the 5% of cases found in the upper and/or lower tail(s) of the distribution.

cross-sectional design A developmental design in which participants from two or more age groups are measured at about the same time. Comparisons are made across age groups to investigate age-related changes in behavior.

data transformation Mathematical operation applied to raw data, such as taking the square root or arcsine of the original scores in a distribution. Often applied to data that violate the assumptions of parametric statistical tests to help them meet those assumptions.

debriefing A session, conducted after an experimental session, in which participants are informed of any deception used and the reasons for the deception.

deception A research technique in which participants are misinformed about the true nature and purpose of a study. Deception is ethical if the researcher can demonstrate that important results cannot be obtained in any other way.

Declaration of Helsinki A declaration on ethical treatment of research participants issued by the World Medical Association in 1964. It stated that the health, welfare, and dignity of research participants be protected by researchers and that research be based on accepted research practices and existing research.

deductive reasoning Reasoning that goes from the general to the specific. Forms the foundation of the rational method of inquiry.

degrees of freedom (*df*) The number of scores that are free to vary in a distribution of a given size having a known mean.

demand characteristics Cues inadvertently provided by the researcher or research context concerning the purposes of a study or the behavior expected from participants.

demonstration A nonexperimental technique in which some phenomenon is demonstrated. No control group is used.

dependent variable The variable measured in a study. Its value is determined by the behavior of the subject and may depend on the value of the independent variable.

descriptive statistics Statistics that allow you to summarize the properties of an entire distribution of scores with just a few numbers.

descriptive theory A theory that simply describes the relationship among variables without attempting to explain the relationship.

direct replication Exactly replicating an experiment. No new variables are included in the replication.

directionality problem A reason not to infer causality from correlational research, stating that the direction of causality is sometimes difficult to determine.

disconfirmational strategy A method of testing a theory that involves conducting research to provide evidence that disconfirms the predictions made by the theory.

discrete trials design A single-subject experimental design in which subjects receive each treatment condition dozens or hundreds of times. Each trial (exposure to a treatment) produces one data point, and data points are averaged across trials to provide stable estimates of behavior.

discriminant analysis Multivariate statistical technique used when you have multiple predictor variables and a categorical criterion variable.

discussion section The section of an APA style manuscript that includes the author's interpretation of the findings of a study and conclusions drawn from the data.

domain The range of situations to which a theory applies. Also called the *scope* of a theory.

double-blind technique Neither the participants in a study nor the person carrying out the study knows at the time of testing which treatment the participant is receiving.

dummy code In a data file, numbers used to stand for category values; for example, 0 = male, 1 = female.

dynamic design An experimental design in which the independent variable is varied continuously over time while monitoring the response of the dependent variable.

effect size The amount by which a given experimental manipulation changes the value of the dependent variable in the population, expressed in standard deviation units.

empirical question A question that can be answered through objective observation.

equivalent time samples design A variation of the time series design in which a treatment is administered repeatedly, with each administration followed by an observation period.

error variance Variability in the value of the dependent variable that is related to extraneous variables and not to the variability in the independent variable.

Ethical Principles of Psychologists and Code of Conduct 2002 A comprehensive document from the American Psychological Association stating the ethical responsibilities of psychologists and researchers.

ethnography A nonquantitative technique used to study and describe the functioning of cultures through a study of social inter actions and expressions between people and groups.

expectancy effect When a researcher's preconceived ideas about how subjects should behave are subtly communicated to subjects and, in turn, affect the subjects' behavior.

experimental group A group of subjects in an experiment that receives a nonzero level of the independent variable.

experimental research Research in which independent variables are manipulated and behavior is measured while extraneous variables are controlled.

experimenter bias When the behavior of the researcher influences the results of a study. Experimenter bias stems from two sources: expectancy effects and uneven treatment of subjects across treatments.

exploratory data analysis (EDA) Examining data for potentially important patterns and relationships, especially through the use of simple graphical techniques and numerical summaries.

external validity The extent to which the results of a study extend beyond the limited sample used in the study.

extraneous variable Any variable that is not systematically manipulated in an experiment but that still may affect the behavior being observed.

face-to-face interview Method of administering a questionnaire that involves face-to-face interaction with the participant. Two types are the structured and unstructured interview.

face validity How well a test appears to measure (judging by its contents) what it was designed to measure. Example: A measure of mathematical ability would have face validity if it contained math problems.

factor analysis Multivariate statistical technique that uses correlations between variables to determine the underlying dimensions (factors) represented by the variables.

factorial design An experimental design in which every level of one independent variable is combined with every level of every other independent variable.

familywise error The likelihood of making at least one Type I error across a number of comparisons.

file drawer phenomenon A problem associated with publication practices and meta-analysis that occurs because results that fail to achieve statistical significance often fail to be published (i.e., get relegated to the researcher's file drawer).

five-number summary A set of five numbers used to summarize the characteristics of a distribution: the minimum, first quartile, median, third quartile, and maximum.

***F* ratio** The test statistic computed when using an analysis of variance. It is the ratio of the between-groups variance to within-groups variance.

frequency distribution A graph or table displaying a set of values or range of values of a variable, together with the frequency of each.

functional explanation An explanation for a phenomenon given in terms of its function, that is, what it accomplishes.

fundamental theory A theory that proposes a new structure or underlying process to explain how variables and constants relate.

generalization Applying a finding beyond the limited situation in which it was observed.

higher-order factorial design Experimental design that includes more than two independent variables (factors).

histogram A graph depicting a frequency distribution in which the frequencies of class intervals are represented by adjacent bars along the scale of measurement.

hypothesis A tentative statement, subject to empirical test, about the expected relationship between variables.

Implicit Association Test (IAT) A popular measure of implicit attitudes that uses responses that are not under direct conscious control.

independent variable The variable that is manipulated in an experiment. Its value is determined by the experimenter, not by the subject.

inferential statistics Statistical procedures used to infer a characteristic of a population based on certain properties of a sample drawn from that population.

informed consent Agreeing to serve as a research participant after being informed about the nature of the research and the participant's rights and responsibilities. The participant typically reads and signs a form specifying the purpose of a study, the methods to be used, requirements for participation, costs and benefits of research participation, that participation is voluntary, and that the participant is free to withdraw from the study at any time without penalty.

institutional animal care and use committee (IACUC) A committee that screens proposals for research using animal subjects and monitors institutional animal-care facilities to ensure compliance with all local, state, and federal laws governing animal care and use.

institutional review board (IRB) A committee that screens proposals for research using human participants for adherence to ethical standards.

interaction When the effect of one independent variable on the dependent variable in a factorial design changes over the levels of another independent variable.

internal validity The extent to which a study evaluates the intended hypotheses.

Internet survey Survey conducted on the Internet, typically by having participants fill out a Web-based questionnaire. Such surveys are subject to potential respondent bias as only those having access to the Internet can respond.

interquartile range A measure of spread in which an ordered distribution of scores is divided into four groups. The score separating the lower 25% is subtracted from the score separating the upper 25%. The resulting difference is divided by 2.

interrater reliability The degree to which multiple observers agree in their classification or quantification of behavior.

interrupted time series design A variation of the time series design in which changes in behavior are charted as a function of time before and after some naturally occurring event.

intersubject replication The behaviors of multiple subjects used in a single-subject design are compared to establish the reliability of results.

interval scale A measurement scale in which the spacing between values along the scale is known. The zero point of an interval scale is arbitrary.

intervention phase Phase of a single-subject, baseline design in which the treatment is introduced and the dependent measure evaluated.

interview *See* ***face-to-face interview.***

intraclass correlation coefficient (r_I) A measure of agreement between observers that can be used when your observations are scaled on an interval or ratio scale of measurement.

intrasubject replication In a single-subject experiment, each treatment is repeated at least once for each subject and behavior is measured. This helps establish the reliability of the results obtained from a single-subject experiment.

introduction The first substantive section of an APA-style manuscript, which includes the rationale for the study, a literature review, and usually a statement of the hypothesis to be tested.

justice An ethical principle in the Belmont Report stating that the researcher and participant should share in the costs and benefits of the research.

latent variable A variable in structural equation modeling that is not directly observable and must be estimated from other measures.

law A relationship that has been substantially verified through empirical test.

lazy writing Flaw in writing, closely related to plagiarism, that involves using too much quoted (albeit properly cited) material in a manuscript.

least-squares regression line Straight line, fit to data, that minimizes the sum of the squared distances between each data point and the line.

linear regression Statistical technique used to determine the straight line that best fits a set of data.

line graph A graph on which data relating the variables are plotted as points connected by lines.

literature review A review of relevant research and theory conducted during the early stages of the research process to identify important variables and accepted methods and to establish a rationale for research hypotheses.

loglinear analysis A nonparametric, multivariate statistical technique used primarily to evaluate data from multifactor research with a nominal dependent variable. It can also be used on interval or ratio data that violate the assumptions of the analysis of variance.

longitudinal design A developmental design in which a single group of subjects is followed over a specified period of time and measured at regular intervals.

mail survey Method of administering a survey that involves mailing questionnaires to participants. Nonresponse bias may be a problem.

main effect The independent effect of one independent variable in a factorial design on the dependent variable. There are as many main effects as there are independent variables.

manipulation check Measures included in an experiment to test the effectiveness of the independent variables.

Mann–Whitney *U* test Nonparametric inferential statistic used to evaluate data from a two-group experiment in which the dependent variable was measured along at least an ordinal scale. It can also be used on interval or ratio data if the data do not meet the assumptions of the *t* test for independent samples.

matched-groups design Between-subjects experimental design in which matched sets of subjects are distributed, at random, one per group across groups of the experiment.

matched-pairs design A two-group matched groups design.

materials subsection A subsection of the method section of an APA-style manuscript in which primarily written materials used in a study (e.g., questionnaires) are described.

mean The arithmetic average of the scores in a distribution. The most frequently reported measure of center.

measure of center A single score, computed from a data set, that represents the general magnitude of the scores in the distribution.

measure of spread A single score, computed from a data set, that represents the amount of variability of the scores in the distribution (i.e., how spread out they are).

mechanistic explanation An explanation for a phenomenon given in terms of a mechanism that is assumed to produce it through an explicit chain of cause and effect.

median The middle score in an ordered distribution.

meta-analysis A statistics-based method of reviewing literature in a field that involves comparing or combining the results of related studies.

method of authority Relying on authoritative sources (e.g., books, journals, scholars) for information.

method section The section of an APA-style manuscript in which the methods used in a study are described in detail.

mixed design An experimental design that includes between-subjects as well as within-subjects factors. Also called a *split-plot design*.

mode The most frequent score in a distribution. The least informative measure of center.

model Specific application of a general theoretical view. The term *model* is sometimes used as a synonym for *theory*.

multiple-baseline design Simultaneously sampling several behaviors in a single-subject, baseline design to provide multiple baselines of behavior. Used if your independent variable produces irreversible changes in the dependent variable.

multiple control group design Single-factor, experimental design that includes two or more control groups.

multiple *R* The correlation between the best linear combination of predictor variables entered into a multiple regression analysis and the dependent variable.

multiple regression Multivariate linear regression analysis used when you have a single criterion variable and multiple predictor variables.

multistage sampling A variant of cluster sampling in which naturally occurring groups of subjects are identified and randomly sampled. Individual subjects are then randomly sampled from the groups chosen.

multivariate analysis of variance (MANOVA) Multivariate analog to the analysis of variance used to analyze data from an experimental design with multiple dependent variables.

multivariate design A research design in which multiple dependent or predictor variables are included.

multivariate strategy A data analysis strategy in which multiple dependent measures are analyzed with a single, multivariate statistical test.

multiway frequency analysis A class of alternatives to ANOVA, MANOVA, or regression analysis for use when you want to measure or manipulate categorical variables.

naturalistic observation Observational research technique in which subjects are observed in their natural environments. The observers remain unobtrusive so that they do not interfere with the natural behaviors of the subjects being observed.

nested design An experimental design with a within-subjects factor in which different levels of one independent variable are included under each level of a between-subjects factor.

nominal scale A measurement scale that involves categorizing cases into two or more distinct categories. This scale yields the least information.

nonequivalent control group design A time series experiment that includes a control group that is not exposed to the experimental treatment.

nonparametric design Experimental research design in which levels of the independent variable are represented by different categories rather than different amounts.

nonparticipant observation An observational research technique in which the observer attends group functions and records observations without participating in the group's activities.

nonrandom sample A specialized sample of subjects used in a study who are not randomly chosen from a population.

nonrefereed journal A journal in which articles do not undergo prepublication editorial review.

nonresponse bias A problem associated with survey research, caused by some participants not returning a questionnaire, resulting in a biased sample.

normal distribution A specific type of frequency distribution in which most scores fall around the middle category. Scores become less frequent as you move from the middle category. Also referred to as a *bell-shaped curve*.

Nuremberg Code An early code of ethical treatment of research participants developed after World War II, resulting from the Nuremberg trials of Nazi war criminal doctors.

Office of Research Integrity (ORI) An office within the U.S. Department of Health and Human Services that oversees the integrity of the research process. The ORI documents and investigates cases involving research fraud.

open-ended item Questionnaire item that allows the subject to fill in a response rather than selecting a response from provided alternatives.

operational definition A definition of a variable in terms of the operations used to measure it.

ordinal scale A measurement scale in which cases are ordered along some dimension (e.g., large, medium, or small). The distances between scale values are unknown.

outliers Values of a variable in a set of data that lie far from the other values.

paper session A meeting at a scientific convention at which the most up-to-date research results are presented. A paper session may involve disseminating data by reading a paper or presenting a poster.

parallel-forms reliability Establishing the reliability of a questionnaire by administering parallel (alternate) forms of the questionnaire repeatedly.

parametric design An experimental design in which the amount of the independent variable is systematically varied across several levels.

parametric statistic A statistic that makes assumptions about the nature of an underlying population (e.g., that scores are normally distributed).

parsimonious explanation An explanation or theory that explains a relationship using relatively few assumptions.

partial correlation Multivariate correlational statistic used to examine the relationship between two variables with the effect of a third variable removed from both of them.

partially open-ended item Questionnaire item that provides participants with response categories but includes an "other" response category with a space for participants to define the category.

participant observation An observational research technique in which a researcher insinuates him- or herself into a group to be studied and participates in the group's activities.

participants subsection A subsection of the method section of an APA-style manuscript used when humans are employed in a study and describing the nature of the sample.

path analysis An application of multiple regression used to develop and test causal models using correlational data.

Pearson product-moment correlation (Pearson *r*) The most popular measure of correlation. Indicates the magnitude and direction of a correlational relationship between variables.

peer review Process of editorial review used by refereed journals. Manuscripts are usually sent out to at least two reviewers who screen the research for quality and importance.

per-comparison error The alpha level for each of any multiple comparisons made among means.

personal communication Information obtained privately from another researcher (e.g., by letter or phone).

phi (φ) coefficient Measure of correlation used when both variables are measured on a dichotomous scale.

physiological measure A measure of a bodily function of subjects in a study (e.g., heart rate).

pie graph Type of graph in which a circle is divided into segments. Each segment represents the proportion or percentage of responses falling in a given category of the dependent variable.

pilot study A small, scaled-down version of a study used to test the validity of experimental procedures and measures.

plagiarism A serious flaw in writing that involves using another person's words or ideas without properly citing the source. *See also* **lazy writing.**

planned comparisons Hypothesis-directed statistical tests made after finding statistical significance with an overall statistical test (such as ANOVA).

point-biserial correlation A variation of the Pearson correlation used when one variable is measured on a dichotomous scale.

population All possible individuals making up a group of interest in a study. For example, all U.S. women constitute a population. A small proportion of the population is selected for inclusion in a study (*see* **sample**).

poster session A way of disseminating research results at a conference, in which a presenter prepares a poster providing information about the research being reported.

power The ability of an experimental design or inferential statistic to detect an effect of a variable when an effect is present.

predictive validity The ability of a measure to predict some future behavior.

pretest–posttest design A research design that involves measuring a dependent variable (pretest), then introducing the treatment, and then measuring the dependent variable a second time (posttest).

primary source A reference source that contains the original, full report of a study. It includes all the details needed to replicate and interpret the study.

procedure subsection The subsection of the method section of an APA-style manuscript that provides a detailed description of the procedures used in a study.

proportionate sampling A variation of stratified sampling in which the proportion of subjects sampled from each stratum is matched to the proportion of subjects in each stratum in the population.

pseudoexplanation An explanation proposed for a phenomenon that simply relabels the phenomenon without really explaining it.

pseudoscience A set of ideas based on theories put forth as scientific when they are not scientific.

PsycARTICLES A computerized source of articles, downloadable in PDF format, that were published in the journals of the American Psychological Association.

PsycINFO A computerized database system that indexes journals and book chapters relevant to psychology and related fields.

***p* value** In a statistical test, the probability, estimated from the data, that an observed difference in sample values arose through sampling error. *p* must be less than or equal to the chosen alpha level for the difference to be statistically significant.

Q-sort methodology A qualitative measurement technique that involves establishing evaluative categories and sorting items into those categories.

qualitative data Data in which the values of a variable differ in kind (quality) rather than in amount.

qualitative theory A theory in which terms are expressed verbally rather than mathematically.

quantitative data Data collected that are represented by numbers that can be analyzed with widely available descriptive and inferential statistics.

quantitative theory A theory in which terms are expressed mathematically rather than verbally.

quasi-experimental design A design resembling an experimental design but using quasi-independent rather than true independent variables.

quasi-independent variable A variable resembling an independent variable in an experiment, but whose levels are not assigned to subjects at random (e.g., the subject's age).

random assignment The process of assigning subjects to experimental treatments randomly.

randomized two-group design A between-subjects design in which subjects are assigned to groups randomly.

random sample A sample drawn from the population such that every member of the population has an equal opportunity to be included in the sample.

range The least informative measure of spread; the difference between the lowest and highest scores in a distribution.

range effects A problem in which a variable being observed reaches an upper limit (ceiling effect) or lower limit (floor effect).

rational method Developing explanations through a process of deductive reasoning.

ratio scale Highest scale of measurement; it has all of the characteristics of an interval scale plus an absolute zero point.

refereed journal A journal whose articles have undergone prepublication editorial review by a panel of experts in the relevant field.

reference section The section of an APA-style manuscript providing an alphabetical list of the bibliographic information for all works cited in a manuscript.

regression weight Value computed in a linear regression analysis that provides the slope of the least squares regression line. *See also* **beta weight.**

reliability Whether a measure or questionnaire produces the same or similar responses with multiple administrations of the same or a similar instrument.

representative sample A sample of subjects in which the characteristics of the population are adequately represented.

resistant measure Statistics that are not strongly affected by the presence of outliers or skewness in the data.

respect for persons An ethical principle in the Belmont Report stating that research participants are free to make their own decisions and that individuals with diminished autonomy must be protected.

restricted item Questionnaire item that provides participants with response alternatives from which the participant selects an answer.

results section The section of an APA-style manuscript that contains a description of the findings of a study. The section normally reports the values of descriptive and inferential statistics obtained.

reversal strategy Running a second baseline phase after the intervention phase in a single-subject, baseline design.

role attitude cue An unintended cue in an experiment that suggests to the participants how they are expected to behave.

role playing Alternative to deceptive research that involves having participants act as though they had been exposed to a certain treatment.

***R*-square** The square of the multiple *R* in a multiple regression analysis. Provides a measure of the amount of variability in the dependent measure accounted for by the best linear combination of predictor variables.

running head A shortened version of the title to a manuscript (no more than 50 characters) that appears on each page of a manuscript.

sample A relatively small number of individuals drawn from a population for inclusion in a study. *See also* **population.**

sampling error The deviation between the characteristics of a sample and a population.

scatter plot A plot used to display correlational data from two measures. Each point represents the two scores provided by each subject, one for each measure, plotted against one another.

science A set of methods used to collect information about phenomena in a particular area of interest and build a reliable base of knowledge about them.

scientific explanation A tentative explanation for a phenomenon, based on objective observation and logic, and subject to empirical test.

scientific method The method of inquiry preferred by scientists. It involves observing phenomena, developing hypotheses, empirically testing the hypotheses, and refining and revising hypotheses.

scientific theory A theory that goes beyond simple hypothesis, deals with verifiable phenomena, and is highly ordered and structured.

scientist A person who adopts the methods of science in his or her quest for knowledge.

secondary source A reference source that summarizes information from a primary source and includes research reviews and theoretical articles.

self-report measure A measure that requires participants to report on their past, present, or future behavior.

semipartial correlation *See* **part correlation.**

simple main effect In a factorial analysis of variance (ANOVA), the effect of one factor at a given level (or combination of levels) of another factor (or factors).

simple random sampling A sampling technique in which every member of a population has an equal chance of being selected for a sample and in which the sampling is done on a purely random basis.

simulation A laboratory research technique in which you attempt to re-create as closely as possible a real-world phenomenon.

single-blind technique The person testing subjects in a study is kept unaware of the hypotheses being tested.

single-subject design An experimental design that focuses on the behavior of an individual subject rather than groups of subjects.

skewed distribution A frequency distribution in which most scores fall into categories above or below the middle category.

sociogram A graphical representation of the pattern of interpersonal relationship choices.

sociometry A nonexperimental research technique involving identifying and measuring interpersonal relationships within a group.

Solomon four-group design An expansion of the pretest–posttest design that includes control groups to evaluate the effects of administering a pretest on your experimental treatment.

Spearman rank-order correlation (rho) A measure of correlation used when variables are measured on at least an ordinal scale.

split-half reliability A method of assessing reliability of a questionnaire using a single administration of the instrument. The questionnaire is split into two parts, and responses from the two parts are correlated.

stability criterion Criterion used to establish when a baseline in a single-subject, baseline design no longer shows any systematic trends. Once the criterion is reached, the subject is placed in the next phase of the experiment.

standard deviation The most frequently reported measure of spread. The square root of the variance.

standard error of estimate A measure of the accuracy of prediction in a linear regression analysis. It is a measure of the distance between the observed data points and the least squares regression line.

standard error of the mean An estimate of the amount of variability in expected sample means across a series of samples. It provides an estimate of the deviation between a sample mean and the underlying population mean.

stemplot A graphical display of a distribution of scores consisting of a column of values (the stems) representing the leftmost digit or digits of the scores and, aligned with each stem, a row of values representing the rightmost digit of each score having that particular stem value.

stratified sampling A sampling technique designed to ensure a representative sample that involves dividing the population into segments (strata) and randomly sampling from each stratum.

strong inference A strategy for testing a theory in which a sequence of research studies is systematically carried out to rule out alternative explanations for a phenomenon.

structural equation modeling (SEM) A variant of path analysis in which variables that are indirectly observed and measured are included in the analysis, allowing you to evaluate relationships involving hypothetical constructs.

subjects subsection A subsection of the method section of an APA-style manuscript in which the nature of the subject sample employed is described. This section is called *subjects* if animals were employed in a study.

systematic replication Conducting a replication of an experiment while adding new variables for investigation.

systematic sampling A sampling technique in which every *k*th element is sampled after a randomly determined start.

systematic variance Variability in the value of the dependent variable that is caused by variation in the independent variable.

tautology *See* **circular explanation.**

telephone survey Method of conducting a survey that involves calling participants on the telephone and asking them questions from a prepared questionnaire.

test–retest reliability A method of assessing the reliability of a questionnaire by administering repeatedly the same or parallel form of a test.

Thesaurus of Psychological Index Terms A thesaurus available in hard copy or in computerized form that is used to help narrow or broaden a search of the psychological literature.

third-variable problem A problem that interferes with drawing causal inferences from correlational results. A third, unmeasured variable affects both measured variables, causing the latter to appear correlated even though neither variable influences the other.

time series design A research design in which behavior of subjects in naturally occurring groups is measured periodically both before and after introduction of a treatment.

title page The first page of an APA-style manuscript, including the running head, title, author name(s) and institutional affiliation(s), and author notes.

treatment A level of an independent variable applied during an experiment. In multifactor designs, a specific combination of the levels of each factor.

***t* test** An inferential statistic used to evaluate the reliability of a difference between two means. Versions exist for between-subjects and within-subjects designs and for evaluating a difference between a sample mean and a population mean.

***t* test for correlated samples** A parametric inferential statistic used to compare the means of two samples in a matched-pairs or a within-subjects design in order to assess the probability that the two samples came from populations having the same mean.

***t* test for independent samples** A parametric inferential statistic used to compare the means of two independent, random samples in order to assess the probability that the two samples came from populations having the same mean.

Type I error Deciding to reject the null hypothesis when, in fact, the null hypothesis is true. Also referred to as an *alpha error*.

Type II error Deciding not to reject the null hypothesis when, in fact, the null hypothesis is false. Also referred to as a *beta error*.

univariate strategy A data analysis strategy in which multiple dependent measures are analyzed independently with separate statistical tests.

unplanned comparison Comparison between means that is not directed by your hypothesis and is made after finding statistical significance with an overall statistical test (such as ANOVA).

validity The extent to which a measuring instrument measures what it was designed to measure.

variable Any quantity or quality that can take on a range of values.

variance A measure of spread. The averaged square deviation from the mean.

volunteer bias Bias in a sample that results from using volunteer participants exclusively.

Wilcoxon signed ranks test A nonparametric statistical test that can be used when the assumptions of the *t* test for correlated samples are seriously violated.

within-subjects design An experimental design in which each subject is exposed to all levels of an independent variable.

***z* test for the difference between two proportions** A parametric inferential statistic used to determine the probability that two independent, random samples came from populations having the same proportion of "successes" (e.g., persons favoring a particular candidate).

REFERENCES

Abbott, B., & Badia, P. (1979). Choice for signaled over unsignaled shock as a function of signal length. *Journal of the Experimental Analysis of Behavior, 32*, 409–417.

Adair, J. G. (1973). *The human subject: The social psychology of the psychological experiment*. Boston: Little, Brown.

Agresti, A., & Finlay, B. (1986). *Statistical methods for the social sciences*. San Francisco: Dullen.

Aguinis, H., & Henle, C. A. (2001). Empirical assessment of the bogus pipeline. *Journal of Applied Social Psychology, 31*, 352–375.

Allport, G. W. (1954). Historical background of modern social psychology. In G. Lindzey (Ed.), *Handbook of social psychology* (Vol. 1, pp. 3–56). Cambridge, MA: Addison-Wesley.

Allport, G. W., & Postman, L. (1945). The basic psychology of rumor. *Transactions of the New York Academy of Sciences, 11*, 61–81.

American Psychological Association. (1973). *Ethical principles in the conduct of research with human participants*. Washington, DC: Author.

American Psychological Association. (1992). Ethical principles of psychologists. *American Psychologist, 45*, 1597–1611.

American Psychological Association. (2001). *Publication manual FAQ*. Washington, DC: Author. Retrieved from http://www.apa.org/journals/faq.html

American Psychological Association. (2002). *Ethical principles of psychologists and code of conduct*. Retrieved from http://www.apa.org/ethics/code2002.html

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Anastasi, A. (1976). Psychological testing (4th ed.). New York: Macmillan.

Anderson, C. A., & Dill, K. E. (2000). Video games and aggressive thoughts, feelings, and behavior in the laboratory and in life. *Journal of Personality and Social Psychology, 78*, 772–790.

Anderson, N. (1968). A simple model for information integration. In R. P. Abelson, E. Aronson, W. J. McGuire, T. M. Newcomb, M. J. Rosenberg, & P. Tannenbaum (Eds.), *Theories of cognitive consistency: A sourcebook* (pp. 731–743). Chicago: Rand McNally.

Anglesea, M. M., Hoch, H., & Taylor, B. A. (2008). Reducing rapid eating in teenagers with autism: Use of a pager prompt. *Journal of Applied Behavior Analysis, 41*, 107–111. doi: 10.1901/jaba.2008.41–107

Applebaum, M. I., & McCall, R. B. (1983). Design and analysis in developmental psychology. In P. H. Mussen & W. Kessen (Eds.), *Handbook of child psychology: Vol. 1. History, theory, and methods* (pp. 415–476). New York: Wiley.

Aronson, E., & Carlsmith, J. M. (1968). Experimentation in social psychology. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (Vol. 1, pp. 1–79). Reading, MA: Addison-Wesley.

Asher, H. B. (1976). Causal modeling. *Sage University paper series on quantitative applications in the social sciences* (Series No. 07003). Beverly Hills, CA: Sage.

Ax, A. F. (1953). The physiological differentiation between fear and anger in humans. *Psychosomatic Medicine, 15*, 432–442.

Badia, P., & Abbott, B. B. (1984). Preference for signaled over unsignaled shock schedules: Ruling out asymmetry and response fixation as factors. *Journal of the Experimental Analysis of Behavior, 41*, 45–52.

Badia, P., & Culbertson, S. (1972). The relative aversiveness of signalled vs. unsignalled escapable and inescapable shock. *Journal of the Experimental Analysis of Behavior, 17*, 463–471.

Badia, P., & Runyon, R. P. (1982). Fundamentals of behavioral research. Reading, MA: Addison-Wesley.

Badia, P., Harsh, J., & Abbott, B. (1979). Choosing between predictable and unpredictable shock conditions: Data and theory. *Psychological Bulletin*, 86, 1107–1131.

Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). Cambridge, England: Cambridge University Press.

Balcetis, E., & Dunning, D. (2007). Cognitive dissonance and the perception of natural environments. *Psychological Science, 18*, 917–921. doi: 10.1111/j.1467-9280.2007.02000.x

Barrett, J., & Jay, P. (2005). Clinical research fraud: A victimless crime? *Applied Clinical Trials, 14*, 44–46.

Baumans, V. (2004). Use of animals in experimental research: An ethical dilemma? *Gene Therapy, 11*, S64–S66.

Baumrind, D. (1964). Some thoughts on the ethics of research: After reading Milgram's "Behavioral study of obedience." *American Psychologist, 26*, 887–896.

Bell, R. (1992). Impure science: Fraud, compromise and political influence in scientific research. New York: Wiley.

Belmont Report (1979). Ethical principles and guidelines for the protection of human subjects of research. Retrieved from http://www.hhs.gov/ohrp/human-subjects/guidance/belmont.htm#xbasic.

Bem, D. J. (1972). Self-perception theory. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 6, pp. 1–62). New York: Academic Press.

Bennett, S., & Bowers, D. (1976). *An introduction to multivariate techniques for social and behavioral sciences*. New York: Wiley.

Berg, B. L. (2009). *Qualitative research methods for the social sciences* (7th ed.). Boston: Allyn and Bacon.

Berman, M. G., Jonides, J., & Kaplan, S. (2008). The cognitive benefits of interacting with nature. *Psychological Science, 19*, 1207–1212. doi: 10.1111/j.1467-9280.2008.02225.x

Bernstein, D. M., Laney, C., Morris, E. K., & Loftus, E. F. (2005). False memories about food can lead to food avoidance. *Social Cognition, 23*, 11–34.

Bethell, C., Fiorillo, J., Lansky, D., Hendryx, M., & Knickman, J. (2004). Online consumer surveys as a methodology for assessing the quality of the United States health care system. *Journal of Medical Internet Research*, 6(1). Retrieved from http://www.jmir.org/2004/.

Blatchley, B., & O'Brien, K. R. (2007). Deceiving the participant: Are we creating the reputational spillover effect? *North American Journal of Psychology*, 9, 519–534.

Block, G. (2003) The moral reasoning of believers in animal rights. *Society and Animals, 11*, 167–180.

Bolt, M., & Myers, D. G. (1983). *Teacher's resource and test manual to accompany social psychology*. New York: McGraw-Hill.

Bordens, K. S. (1984). The effects of likelihood of conviction, threatened punishment, and assumed role on mock plea bargain decisions. *Basic and Applied Social Psychology, 5*, 59–74.

Bordens, K. S., & Horowitz, I. A. (1986). Prejudicial joinder of multiple offenses: The relative effects of cognitive processing and criminal schemata. *Basic and Applied Social Psychology*, 7, 243–258.

Bosnjak, M., Neubarth, W., Couper, M. P., Bandilla, W., & Kaczmirek, L. (2008). Prenotification in Web-based access panel surveys: The influence of mobile text messaging versus e-mail on response rates and sample composition. *Social Science Computer Review, 26*, 213–222.

Boynton, P. (2003). "I'm just a girl who can't say no"?: Women, consent, and sex research. *Journal of Sex and Marital Therapy, 29*, 23–32.

Braithwaite, R. B. (1953). *Scientific explanation*. New York: Harper & Row.

Bray, J. H., & Maxwell, S. E. (1982). Analyzing and interpreting significant MANOVAs. *Review of Educational Research, 52*, 340–367.

Broad, W., & Wade, N. (1983). *Betrayers of the truth*. New York: Simon & Schuster.

Broca, P. P. (1861). Loss of speech, chronic softening and partial destruction of the anterior left lobe of the brain. Retrieved from http://psychclassics.yorku.ca/Broca/perte-e.htm.

Brody, J. L., Gluck, J. P., & Aragon, A. S. (2000). Participants' understanding of the process of psychological research: Debriefing. *Ethics and Behavior*, *10*, 13–25.

Brown, R. (1965). *Social psychology*. New York: Free Press.

Brown, S. R. (1996). Q methodology and qualitative research. *Qualitative Health Research*, *6*, 561–567. Retrieved from http://www.rz.unibw-muenchen.de/~p41bsmk/qmethod/srbqhc.htm.

Bruning, J. L., & Kintz, B. L. (1987). *Computational handbook of statistics* (3rd ed.). Glenview, IL: Scott, Foresman.

Butler, B. E., & Petrulis, J. (1999). Some further observations concerning Cyril Burt. *British Journal of Psychology*, *90*, 155–160.

Cameron, L., & Rutland, A. (2006). Extended contact through story reading in school: Reducing children's prejudice toward the disabled. *Journal of Social Issues*, *62*, 469–488.

Campbell, D. T. (1969). Prospective: Artifact and control. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research* (pp. 351–382). New York: Academic Press.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Carnahan, T., & McFarland, S. (2007). Revisiting the Stanford Prison Experiment: Could participant self-selection have led to the cruelty? *Personality and Social Psychology Bulletin*, *33*, 603–614.

Carroll, R. T. (2006). Pseudoscience. Retrieved from http://skepdic.com/pseudosc.html.

Ceci, S. J., Bruck, M., & Loftus, E. F. (1998). On the ethics of memory implantation research. *Applied Cognitive Psychology*, *12*, 230–240.

Chabris, C. F., & Glickman, M. E. (2006). Sex differences in intellectual performance: Analysis of a large cohort of competitive chess players. *Psychological Science*, *17*, 1040–1046.

Chang, D. F., & Sue, S. (2003). The effects of race and problem type on teachers' assessments of student behavior. *Journal of Consulting and Clinical Psychology*, *71*, 235–242.

Chassan, J. B. (1967). *Research design in clinical psychology and psychiatry*. New York: Appleton-Century-Crofts.

Chomsky, N. (1965). *Aspects of a theory of syntax*. Cambridge, MA: MIT Press.

Church, A. H. (1993). Estimating the effects of incentives on mail survey return rates: A meta-analysis. *Public Opinion Quarterly*, *57*, 62–79.

Cialdini, R. B. (1994). A full-cycle approach to social psychology. In G. G. Brannigan & M. R. Merrens (Eds.), *The social psychologists: Research adventures* (pp. 52–72). New York: McGraw-Hill.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. M., Bankert, E., & Cooper, J. A. (2005). History and ethics. Retrieved from https://www.citiprogram.org/members/courseandexam/moduletext.asp?strKeyID=0332B55E-959E-49B2-A091-4B344164DA45-1045638.

Cohen, R. J., & Swerdlik, M. E. (2010). *Psychological testing and assessment: An introduction to tests and measures* (7th ed.). Boston: McGraw-Hill.

Coker, R. (2007). Distinguishing science from pseudoscience. Retrieved from https://webspace.utexas.edu/cokerwr/www/index.html/distinguish.htm.

Conrad, E., & Maul, T. (1981). *Introduction to experimental psychology*. New York: Wiley.

Cooper, H. M., & Rosenthal, R. (1980). Statistical versus traditional methods for summarizing research findings. *Psychological Bulletin*, *87*, 442–449.

Cooperman, E. (1980). Voluntary subjects' participation in research: Cognitive style as a possible biasing factor. *Perceptual and Motor Skills*, *50*, 542.

Cornell University Library. (2000). *Distinguishing scholarly journals from other periodicals*. Retrieved from http://www.library.cornell.edu/okuref/research/skill20.html#scholarly.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology*, *83*, 1314–1329.

Crano, W. D., & Brewer, M. B. (1986). *Principles and methods of social research*. Boston: Allyn and Bacon.

Crews, F. (1980). *The Random House handbook* (3rd ed.). New York: Random House.

Davidson, B., Worall, L., & Hickson, L. (2003). Identifying communication activities of older people with aphasia: Evidence from naturalistic observation. *Aphasiology*, *17*, 243–264.

Davidson, P. R., & Parker, K. C. H. (2001). Eye movement desensitization and reprocessing

(EMDR): A meta-analysis. *Journal of Consulting and Clinical Psychology*, *69*, 216–316.

Davis, A. J. (1984). Sex-differentiated bias in nonsexist picture books. *Sex Roles: A Journal of Research*, *11*, 1–16.

Davis, M. H., Mitchell, K.V., Hall, J. A., Lothert, J., Snapp, T., & Meyer, M. (1999). Empathy, expectations, and situational preferences: Personality influences on the decision to participate in volunteer helping behaviors. *Journal of Personality*, *67*, 469–503.

De Beuckelear A., & Lievens, F. (2009). Measurement equivalence of paper-and-pencil and Internet organisational surveys: A large scale examination in 16 countries. *Applied Psychology: An International Review*, *58*, 336–361. doi: 10.1111/j.1464-0597.2008.00350.x

DeSantis, A. D. (2003). A couple of White guys sitting around talking: The collective rationalization of cigar smokers. *Journal of Contemporary Ethnology*, *32*, 432–466.

Dewsbury, D. A. (1978). *Comparative animal behavior*. New York: McGraw-Hill.

DeWall, N. C. & Baumeister, R. F. (2007). From terror to joy: Automatic tuning to positive affective information following mortality salience. *Psychological Science*, *18*, 984–990. doi: 10.1111/j.1467-9280.2007.02013.x

Dillman, D. A. (2000). *Mail and Internet surveys: The tailored design method* (2nd ed.). New York: Wiley.

DiNitto, D. M, Busch-Armendariz, N. B., Bender, K., Woo, H., Tackett-Gibson, M., & Dyer, J. (2009). Testing telephone and web surveys for studying men's sexual assault penetration behaviors. *Journal of Interpersonal Violence*, *23*, 1483–1493. doi: 10.1177/0886260508314341

Drews, F. A., Pasupathi, M., & Strayer, D. L. (2008). Passenger and cell conversations in simulated driving. *Journal of Experimental Psychology: Applied*, *14*, 392–400. doi: 10.1037/a0013119

Ebbinghaus, H. E. (1964). *Memory: A contribution to experimental psychology*. New York: Dover. (Original work published 1885)

Edwards, A. L. (1953). *Techniques of attitude scale construction*. New York: Appleton-Century-Crofts.

Edwards, A. L. (1985). *Experimental design in psychological research* (5th ed.). New York: Harper & Row.

Ellis, C. (1993). "There are survivors": Telling a story of sudden death. *Sociological Quarterly*, *34*, 711–730.

Epley, N., & Huff, C. (1998). Suspicion, affective response, and educational benefi t as a result of deception in psychology research. *Personality and Social Psychology Bulletin*, *24*, 759–768.

Fancher, R. E. (1979). *Pioneers of psychology*. New York: Norton.

Fancher, R. E. (1985). *The intelligence men: Makers of the IQ controversy*. New York: Norton.

Feild, H. S., & Barnett, N. J. (1978). Students vs. "real" people as jurors. *Journal of Social Psychology*, *104*, 287–293.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.

Fischer, H., Anderson, J. L. R., Furmark, T., Wik, G., & Fredrikson, M. (2002). Right-sided human prefrontal brain activation during acquisition of conditioned fear. *Emotion*, 233–241.

Fishbein, M., & Ajzen, I. (1975). Belief, attitude, intention and behavior: An introduction to theory and research. Reading, MA: Addison-Wesley.

Fisher, W. W., Kelley, M. E., & Lomas, J. E. (2003). Visual aids and structured criteria for improving visual inspection and interpretation of single-case designs. *Journal of Applied Behavior Analysis*, *36*, 387–406.

Fiske, D. W., & Fogg, L. (1990). But the reviewers are making different criticisms of my paper! *American Psychologist*, *45*, 591–598.

Fiske, S. T. (2009). Institutional review boards: From bane to benefit. *Perspectives on Psychological Science*, *4*, 30–31. doi: 10.1111/j.1745-6924.2009.01085.x

Fleming, C., & Bowden, M. (2009). Web-based *surveys* as an alternative to traditional mail methods. *Journal of Environmental Management*, *90*, 284–292. doi:10.1016/j.jenvman.2007.09.011

Florian, V., Mikulincer, M., & Harschberger, H. (2002). The anxiety-buffering function of close relationships: Evidence that relationship commitment acts as a terror management mechanism. *Journal of Personality and Social Psychology*, *82*, 527–542.

Forscher, B. K. (1963). Chaos in the brickyard. *Science*, *42*, 339.

Foundation for Biomedical Research (2005). Poll shows a majority of Americans favor animal research. Retrieved from http://www.fbresearch.org/Journalists/Releases/Polls/HartPoll_4_15_05.htm.

Freedman, J. L. (1969). Role playing: Psychology by consensus. *Journal of Personality and Social Psychology*, *13*, 107–114.

Fry, D. P. (1992). "Respect for the rights of others is peace": Learning aggression versus nonaggression among the Zapotec. *American Anthropologist*, *94*, 621–639.

Gabriel, U., & Banse, R. (2006). Helping behavior as a subtle measure of discrimination against lesbians and gay men: German data and a comparison across countries. *Journal of Applied Social Psychology*, *36*, 690–707.

Gaither, G. A., Sellbom, M., & Meier, B. P. (2003). The effect of stimulus content on volunteering for sexual interest research among college students. *Journal of Sex Research*, *40*, 240–248.

Gamson, W. A., Fireman, B., & Rytina, S. (1982). *Encounters with unjust authority*. Homewood, IL: Dorsey Press.

Garcia, J., & Koelling, R. A. (1966). Relation of cue to consequences in avoidance learning. *Psychonomic Science*, *4*, 123–124.

Garson, D. G. (2006). Structural equation modeling. Retrieved from http://www2.chass.ncsu.edu/garson/pa765/structur.htm.

GazetteOnline (2005). Skorton testifies in D.C. about lab vandalism. Retrieved from http://www.gazetteonline.com/2005/05/18/Home/News/skortontestimony.prt.

Geggie, D. (2001). A survey of newly appointed consultants' attitudes towards fraud in research. *Journal of Medical Ethics*, *27*, 344–346.

Gibbon, J. (1977). Scalar expectancy theory and Weber's law in animal timing. *Psychological Review*, *84*, 279–325.

Glass, G. V. (1978). In defense of generalization. *The Behavioral and Brain Sciences*, *1*(3), 394–395.

Gold, P. E. (1987). Sweet memories. *American Scientist*, *75*, 151–155.

Goldiamond, I. (1965). Stuttering and fluency as manipulable operant response classes. In L. Krasner & L. P. Ullman (Eds.), *Research in behavior modification* (pp. 106–156). New York: Holt, Rinehart & Winston.

Goldstein, I. (2003). Sexual dysfunction after hysterectomy. Retrieved from http://www.bumc.bu.edu/Dept/Contentaspx?DepartmentID=371&PageID=7309.

Goldstein, J. H., Rosnow, R. L., Goodstadt, B. E., & Suls, J. E. (1972). The good subject in verbal operant conditioning research. *Journal of Experimental Research in Personality*, *28*, 29–33.

Gottman, J. M., & Roy, A. K. (2008). *Sequential analysis: A guide for behavioral researchers*. New York: Cambridge University Press.

Gravetter, F. J., & Wallnau, L. B. (2010). *Statistics for the behavioral sciences* (8th ed.). Belmont, CA: Wadsworth.

Greenberg, B. S. (1980). *Life on television: Current analyses of U.S. TV drama*. Norwood, NJ: Ablex.

Greene, E., & Loftus, E. F. (1984). What's in the news? The influence of well-publicized news events on psychological research and courtroom trials. *Basic and Applied Social Psychology*, *5*, 211–221.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*, 1464–1480.

Grice, G. R. (1966). Dependence of empirical laws upon the source of experimental variation. *Psychological Bulletin*, *66*, 488–498.

Hall, D. (1979). *Writing well*. Boston: Little, Brown.

Hall, R. V., Lund, D., & Jackson, D. (1968). Effects of teacher attention on study behavior. *Journal of Applied Behavior Analysis*, *1*, 1–12.

Hamby, S., Sugarman, D. B., & Boney-McCoy, S. (2006). Does questionnaire format impact reported partner violence rates? An experimental study. *Violence and Victims*, *21*, 507–518.

Haney, C., Banks, C., & Zimbardo, P. (1973). Interpersonal dynamics in a simulated prison. *International Journal of Criminology and Penology*, *1*, 69–87.

Harari, H., Harari, O., & White, R. V. (1985). The reaction to rape by American male by-standers. *Journal of Social Psychology*, *125*, 653.

Helmstetter, F. J., & Fanselow, M. S. (1987). Strain differences in reversal of conditional analgesia by opioid antagonists. *Behavioral Neuroscience*, *101*, 735–737.

Hempel, C. G. (1966). *Philosophy of natural science*. Englewood Cliffs, NJ: Prentice Hall.

Herrmann, D., & Yoder, C. (1998). The potential effects of the implanted memory paradigm on child subjects. *Applied Cognitive Psychology*, *12*, 198–206.

Herrnstein, R. J. (1970). On the law of effect. *Journal of the Experimental Analysis of Behavior*, *13*, 243–266.

Herrnstein, R. J., & Prelec, D. (1992). Melioration. In G. Loewenstein & J. Elster (Eds.), *Choice over time* (pp. 235–263). New York: Russell Sage.

Hershberger, S. L., Marcoulides, G. A., & Parramore, M. M. (2003). Structural equation modeling: An introduction. Retrieved from www.loc.gov/catdir/samples/cam033/2002035067.pdf.

Hertwig, R., & Ortmann, A. (2008). Deception in experiments: Revisiting the arguments in its defense. *Ethics and Behavior*, *18*, 59–92. doi: 10.1080/10508420701712990

Hess, D. W., Marwitz, J., & Kreutzer, J. (2003). Neuropsychological impairments in SCI. *Rehabilitation Psychology*, *48*(3).

Higbee, K. L., Millard, R. J., & Folkman, J. R. (1982). Social psychology research during the 1970s: Predominance of experimentation on college students. *Personality and Social Psychology Bulletin*, *8*, 182–183.

Hite, S. (1976). *The Hite report: A nationwide study on female sexuality*. New York: Macmillan.

Hite, S. (1983). *The Hite report on male sexuality*. New York: Ballantine Books.

Holmes, D. S. (1976a). Debriefing after psychological experiments I: Effectiveness of postdeception dehoaxing. *American Psychologist*, *31*, 858–867.

Holmes, D. S. (1976b). Debriefing after psychological experiments II: Effectiveness of postdeception desensitizing. *American Psychologist*, *31*, 868–875.

Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Reading, MA: Addison-Wesley.

Hooke, R. (1983). *How to tell the liars from the statisticians*. New York: Dekker.

Horowitz, I. A. (1969). The effects of volunteering, fear arousal, and number of communications on attitude change. *Journal of Personality and Social Psychology*, *11*, 34–37.

Horowitz, I. A. (1985). The effects of jury nullification instructions on verdicts and jury functioning in criminal trials. *Law and Human Behavior*, *9*, 25–36.

Horowitz, I. A., & Bordens, K. S. (1988). The effects of outlier presence, plaintiff population size, and aggregation of plaintiffs on simulated jury decisions. *Law and Human Behavior*, *13*, 209–229.

Horowitz, I. A., & Rothschild, B. H. (1970). Conformity as a function of deception and role playing. *Journal of Personality and Social Psychology*, *14*, 224–226.

Horowitz, I. A., Bordens, K. S., & Feldman, M. S. (1980). A comparison of verdicts obtained in severed and joined criminal trials. *Journal of Applied Social Psychology*, *10*, 444–456.

Huang, H-M. (2005). Do print and Web surveys provide the same results? *Computers in Human Behavior*, *22*, 334–350.

Huck, S. W., & Sandler, H. M. (1979). *Rival hypotheses: Alternative explanations of data based conclusions*. New York: Harper & Row.

Hudson, J. M., & Bruckman, A. (2004). "Go away": Participant objections to being studied and the ethics of chatroom research. *The Information Society*, *20*, 127–139.

Hunter, J. E. (1987). Multiple dependent variables in program evaluation. In M. M. Mark & R. L. Shotland (Eds.), *Multiple methods*. San Francisco: Jossey-Bass.

Hunter, J. E., & Gerbing, D. W. (1982). Unidimensional measurement, second-order factor analysis and causal models. *Research in Organizational Behavior*, *4*, 267–320.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Institute for Scientific Information. (1988). SSCI journal citation reports: A bibliometric analysis of social science journals in the ISI data base. *Social Sciences Citation Index*, *6*.

James, J. M., & Bolstein, R. (1990). The effect of monetary incentives and follow-up mailings on the response rat and response quality in mail surveys. *Public Opinion Quarterly*, *54*, 346–361.

James, J. M., & Bolstein, R. (1992). Large monetary incentives and their effect on mail survey response rates. *Public Opinion Quarterly*, 56, 442–453

Janis, I., & Mann, L. (1965). Effectiveness of emotional role playing in modifying smoking habits and attitudes. *Journal of Experimental Research in Personality*, *1*, 84–90.

Jones, R. A. (1994). The ethics of research in cyberspace. *Internet Research*, *4*, 30–35.

Joynson, R. B. (1989). *The Burt affair*. London: Routledge.

Kalichman, M., & Friedman, P. (1992). A pilot study of biomedical trainees perceptions concerning research ethics. *Academic Medicine*, *67*, 769–775.

Kanuk, L., & Berenson, C. (1975). Mail surveys and response rates: A literature review. *Journal of Marketing Research*, *12*, 440–453.

Kardes, F. (1996). In defense of experimental consumer psychology. *Journal of Consumer Psychology*, *5*, 279–296.

Kassam, K. S., Gilbert, D. T., Swencionis, J. K., & Wilson, T. D. (2009). Misconceptions of memory: The Scooter Libby effect. *Psychological Science*, *18*, 551–552. doi: 10.1111/j.1467-9280.2009.02334.x

Katz, J. (1972). *Experimentation with human beings*. New York: Russell Sage Foundation.

Kazdin, A. E. (1976). Statistical analyses for single-case experimental designs. In M. Hersen & D. H. Barlow (Eds.), *Single-case experimental designs: Strategies for studying behavior change* (pp. 265–316). New York: Pergamon Press.

Kazdin, A. E. (1978). Methodological and interpretive problems of single-case experimental designs. *Journal of Consulting and Clinical Psychology*, *46*, 629–642.

Kelly, G. (1963). *A theory of personality: The psychology of personal constructs*. New York: Norton.

Kelman, H. C. (1967). The use of human subjects: The problem of deception in social psychological experiments. *Psychological Bulletin*, *67*, 1–11.

Keppel, G. (1973). *Design and analysis: A researcher's handbook*. Englewood Cliffs, NJ: Prentice Hall.

Keppel, G. (1982). *Design and analysis: A researcher's handbook* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Key, W. B. (1973). *Subliminal seduction*. New York: Signet.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Krantz, J. H., & Dalal, R. (2000). Validity of Web-based research. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 35–57). San Diego, CA: Academic Press.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Thousand Oaks, CA: Sage.

Kruse, C. R. (1999). Gender, views of nature, and support for animal rights. *Society and Animals*, *7*, 179–198.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*, 480–498.

Landy, E., & Aronson, E. (1969). The influence of the character of the criminal and his victim on the decisions of simulated jurors. *Journal of Experimental Social Psychology*, *5*, 141–152.

Latané, B. (1981). The psychology of social impact. *American Psychologist*, *36*, 343–356.

Leaton, R. N., & Borszcz, G. S. (1985). Potentiated startle: Its relation to freezing and shock intensity in rats. *Journal of Experimental Psychology: Animal Behavior Processes*, *11*, 421–428.

Leggett, G., Mead, C. D., & Charvat, W. (1978). *Prentice Hall handbook for writers* (7th ed.). Englewood Cliffs, NJ: Prentice Hall.

Levine, M. S. (1977). Canonical analysis and factor comparison. *Sage University paper series on quantitative applications in the social sciences* (Series No. 07-006). Beverly Hills, CA: Sage.

Lewis, J. E. (2008). Dream reports of animal rights activists. *Dreaming*, *18*, 181–200. doi: 10.1037/a0013393

Lilienfeld, S. O., Lynn, S. J., & Lohr, J. M. (2003). Science and pseudoscience in clinical psychology: Initial thoughts, reflections, and considerations. In S. O. Lilienfeld, S. J. Lynn, & J. M. Lohr (Eds.). *Science and pseudoscience in clinical psychology* (pp. 1–14). New York: Guilford Press.

Lilienfeld, S. O. (2005). The 10 commandments of helping students distinguish science from pseudoscience in psychology. *APS Observer*, *18*. Retrieved from http://www.psychologicalscience.org/observer/getArticle.cfm?id=1843.

Lindsey, D. (1978). *The scientific publication system in social science*. San Francisco: Jossey-Bass.

Link, M. W., & Mokdad, A. (2005). Effects of survey mode on self-reports of alcohol consumption: A comparison of mail, Web, and telephone approaches. *Journal of Studies on Alcohol*, 66, 239–245.

Loftus, E. F. (1979). *Eyewitness testimony*. Cambridge, MA: Harvard University Press.

Longino, H. E. (1990). *Science as social knowledge*. Princeton, NJ: Princeton University Press.

Lönnqvist, J-E., Paunonen, S., Verkaslo, M., Leikas, S., Tuulio-Henrikkson, A., & Lönnqvist, J. (2006). Personality characteristics of research volunteers. *European Journal of Personality*, *21*, 1017–1030. doi: 10.1002/per.655

Lord, F. M. (1953). On the statistical treatment of football numbers. *American Psychologist*, *8*, 750–751.

Lorenz, K. (1950). The comparative method in studying innate behavior patterns. *Symposium of the Society for Experimental Biology*, *4*, 221–268.

Macaulay, D. (1979). *Motel of the mysteries*. Boston: Houghton Mifflin.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research*, *1*, 161–175.

Mandel, F. S., Weiner, M., Kaplan, S., Pelcovitz, D., & Labruna, V. (2000). An examination of bias in volunteer subject selection: Findings from an in-depth child abuse study. *Journal of Traumatic Stress*, *13*, 77–88.

Manolov, R., Solanas, A., Bulte, I., & Onghena P. (2010). Data-driven-specific robustness and power of randomization tests for ABAB designs. *The Journal of Experimental Education*, *78*, 191–214. doi: 10.1080/00220970903292827

Mans, G., & Stream, C. (2006). Relationship between news media coverage of medical research and academic medical centers and people volunteering for clinical trials. *Public Relations Review*, *32*, 196–198.

Marcus, B., & Schütz, A. (2005). Who are the people reluctant to participate in research? Correlates of four different types of nonresponse as inferred from self- and observer ratings. *Journal of Personality*, *73*, 959–964.

Martin, E. (1985). *Doing psychology experiments* (2nd ed.). Monterey, CA: Brooks/Cole.

Matfield, M. (2002). Animal experimentation: The continuing debate. *Nature Reviews*, *1*, 149–152.

Matyas, T. A., & Greenwood, K. M. (1990). Visual analysis of single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis*, *23*, 341–351.

Mayo, C., & LaFrance, M. (1977). *Evaluating research in social psychology*. Monterey, CA: Brooks/Cole.

McEnvoy, S. P., Stevenson, M. R., McCartt, A. T., et al. (2005). Role of mobile phones in motor vehicle crashes resulting in hospital attendance: A case-crossover study. *BMJ* (published 12 July 2005), doi:10.1136/bmj.38537.397512.55.

McFarland, C., Cheam, A., & Buehler, R. (2007). The perseverance effect in the debriefing paradigm: Replication and extension. *Journal of Experimental Social Psychology*, *43*, 223–240.

McFarland, S. (1981). Effects of question order on survey responses. *Public Opinion Quarterly*, *48*, 208–215.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30–46.

McNemar, Q. (1946). Opinion-attitude methodology. *Psychological Bulletin*, *43*, 289–374.

Michael, R. T., Gagnon, J. H., Laumann, E. O., & Kolata, G. (1994). *Sex in America: A definitive survey*. Boston: Little, Brown.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, *67*, 371–378.

Milgram, S. (1974). *Obedience to authority*. New York: Harper & Row.

Milton, R. C. (1964). Extended tables for the Mann–Whitney (Wilcoxon) two-sample test. *Journal of the American Statistical Association*, *59*, 925–934.

Mitchell, S. K. (1979). Interobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychological Bulletin*, *86*, 376–390.

Montee, B. B., Miltenberger, R. G., & Wittrock, D. (1995). An experimental analysis of facilitated communication. *Journal of Applied Behavior Analysis*, *28*, 189–200.

Mook, D. G. (1983). In defense of external validity. *American Psychologist*, *38*, 379–387.

Moore, D. S., & McCabe, G. P. (2006). *Introduction to the practice of statistics* (5th ed.). New York: Freeman.

Moser, C. A., & Kalton, G. (1972). *Survey methods in social investigation*. New York: Basic Books.

Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression*. New York: Addison-Wesley.

Myers, D. G. (1999). *Social psychology* (6th ed.). New York: McGraw-Hill.

National Research Council. (1996). *Guide for the care and use of laboratory animals*. Washington, DC: National Academy Press.

Neisser, U. (1976). *Cognition and reality: Principles and implications for cognitive psychology*. San Francisco: Freeman.

Nerb, J., & Spada, H. (2001). Evaluation of environmental problems: A coherence model of cognition and emotion. *Cognition and Emotion*, *15*, 521–551.

Nirenberg, T. D. Wincze, J. P., Bansal, S., Liepman, M. R., Engle-Friedman, M., & Begin, A. (1991). Volunteer bias in a study of male alcoholics' sexual behavior. *Archives of Sexual Behavior*, *20*, 371–379.

Nunnally, J. C. (1967). *Psychometric theory*. New York: McGraw-Hill.

O'Brien, R. G., & Kaiser, M. K. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin*, *97*, 316–333.

Occam's Razor (n.d.). Retrieved from http://www.2think.org/occams_razor.shtml.

Oczak, M., & Niedźwieńska, A. (2007). Debriefing in deceptive research: A proposed new procedure. *Journal of Empirical Research on Research Ethics*, *2*, 49–59. doi:10.1525/jer.2007.2.3.49

Oei, A., & Hartley, L. R. (2005). The effects of caffeine and expectancy on attention and memory. *Human Pharmacology*, *20*, 193–202.

Office of Research Integrity (1995). Consequences of whistleblowing for the whistleblower in misconduct in science cases. Retrieved from http://ori.hhs.gov/documents/final.pdf.

Office of Research Integrity (2003). Handling misconduct: Whistleblowers. Retrieved from http://ori.dhhs.gov/html/misconduct/whistleblowers.asp.

Office of Research Integrity (2006). Case summaries. Retrieved from http://ori.dhhs.gov/documents/newsletters/vol14_no2.pdf.

Office of Research Integrity (2007). 2007 annual report. Retrieved from http://ori.dhhs.gov/documents/annual_reports/ori_annual_report_2007.pdf.

Office of Research Integrity (2009). Handling misconduct-complaintant. Retrieved from http://ori.hhs.gov/misconduct/whistleblowers.shtml.

Ogloff, J. R. P., & Vidmar, N. (1994). The impact of pretrial publicity on jurors: A study to compare the relative effects of television and print media in a child sex abuse case. *Law and Human Behavior*, *18*, 507–525.

Orbell, S., & Hagger, M. (2006). Temporal framing and the decision to take part in type 2 diabetes screening: Effects of individual differences in consideration of future consequences on persuasion. *Health Psychology*, *25*, 537–548.

Orne, M. T. (1962). On the social psychology of the psychological experiment with particular reference to demand characteristics and their implications. *American Psychologist*, *17*, 776–783.

Ornstein, P. A., & Gordon, B. N. (1998). Risk versus rewards of applied research with children: Comments on "The potential effects of the implanted memory paradigm on child participants" by Douglas Herrmann and Carol Yoder. *Applied Cognitive Psychology*, *12*, 241–244.

Pagano, R. R. (2010). *Understanding statistics in the behavioral sciences* (9th ed.). Belmont, CA: Wadsworth.

Palya, W. L., Walter, D., Kessel, R., & Lucke, R. (1996). Investigating behavioral dynamics with a fixed-time extinction schedule and linear analysis. *Journal of the Experimental Analysis of Behavior*, *66*, 391–409.

Pearson, E. S., & Hartley, H. O. (Eds.). (1966). *Biometrika: Tables for statisticians* (Vol. 1, 3rd ed.). London: Cambridge University Press.

Peplau, L. A., & Conrad, E. (1989). Beyond nonsexist research: The perils of feminist methods in psychology. *Psychology of Women Quarterly*, *13*, 379–400.

Peters, D. P., & Ceci, S. J. (1982). Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences*, *5*, 187–255.

Peterson, L. R., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, *58*, 193–198.

Piaget, J. (1952). *The origins of intelligence in children*. New York: Norton.

Pittinger, D. J. (2002). Deception in research: Distinctions and solutions from the perspective of utilitarianism. *Ethics and Behavior*, *12*, 117–142.

Pittenger, D. J. (2003). Internet research: An opportunity to revisit classic ethical problems in behavioral research. *Ethics and Behavior*, *13*, 45–60.

Platt, J. R. (1964). Strong inference. *Science*, *146*, 347–353.

Plous, S. (1996). Attitudes toward the use of animals in psychological research and education: Results from a national survey of psychologists. *American Psychologist*, *51*, 1167–1180.

Plous, S. (1998). Signs of change within the animal rights movement: Results from a follow-up survey of activists. *Journal of Comparative Psychology*, *112*, 48–54.

Powers, W. T. (1978). Quantitative analysis of purposive systems: Some spadework at the foundations of scientific psychology. *Psychological Review*, *85*, 417–435.

Probst, T. (2003). Exploring employee outcomes of organizational restructuring. *Group and Organization Management*, *28*, 416–439.

Redelmeier, D. A., & Tibshirani, R. J. (1997). Association between cellular telephone calls and motor vehicle collisions. *New England Journal of Medicine*, *336*, 453–458.

Reed, D. D., Critchfield, T. S., & Martins, B. K. (2006). The generalized matching law in elite sport competition: Football play calling as operant choice. *Journal of Applied Behavior Analysis*, *39*, 281–297. doi: 10.1901/jaba.2006.146-05

Renfrey, G., & Spates, C. R. (1994). Eye movement desensitization: A partial dismantling study. *Journal of Behavioral Therapy and Experimental Psychiatry*, *25*, 231–239.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokosy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Resnick, J. H., & Schwartz, T. (1973). Ethical standards as an independent variable in psychological research. *American Psychologist*, *28*, 134–139.

Reynolds, G. S. (1961). Attention in the pigeon. *Journal of the Experimental Analysis of Behavior*, *4*, 203–208.

Rhodes, J. C., Kjerulff, K. H., Langenberg, P. W., & Guzinski, G. M. (1999). Hysterectomy and sexual functioning. *Journal of the American Medical Association*, *282*, 1934–1941.

Riva, G., Teruzzi, T., & Anolli, L. (2003). The use of the Internet in psychological research: Comparison of online and offline questionnaires. *CyberPsychology and Behavior*, *6*, 73–80.

Roberts, J. V. (1985). The attitude–memory relationship after 40 years: A meta-analysis of the literature. *Basic and Applied Social Psychology*, *6*, 221–242.

Rogers, T. B. (1995). *The psychological testing enterprise: An introduction*. Pacific Grove, CA: Brooks/Cole.

Roscoe, J. T. (1975). *Fundamental statistics for the behavioral sciences* (2nd ed.). New York: Holt, Rinehart & Winston.

Rosenthal, R. (1976). *Experimenter effects in behavioral research* (enlarged ed.). New York: Irvington.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.

Rosenthal, R. (1984). Meta-analytic procedures for social research. *Applied Social Research Methods* (Vol. 6). Beverly Hills, CA: Sage.

Rosenthal, R., & Rosnow, R. L. (1975). *The volunteer subject*. New York: Wiley.

Rosnow, R. L., & Rosnow, M. (1986). *Writing psychology papers*. Monterey, CA: Brooks/Cole.

Ross, L., Lepper, M. R., & Hubbard, M. (1975). Perseverance in self-perception and social perception: Biased attributional processes in debriefing paradigms. *Journal of Personality and Social Psychology*, *32*, 880–892.

Sagar, H. A., & Schofield, J. W. (1980). Racial and behavioral cues in Black and White children's perceptions of ambiguously aggressive acts. *Journal of Personality and Social Psychology*, *19*, 590–598.

Saini, J., Kuczynski, E., Gretz, H. F., III, & Sills, E. S. (2002). Supracervical hysterectomy versus total abdominal hysterectomy: Perceived effects on sexual function. *BMC Women's Health*, *1*. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=65528.

Salmivalli, C. A., Kaukiainen, A., & Lagerspetz, K. (2000). Aggression and sociometric status among peers: Do gender and type of aggression matter? *Scandinavian Journal of Psychology, 41*, 17–24.

Saunders, D. R. (1980). Definition of Stroop interference in volunteers and nonvolunteers. *Perceptual and Motor Skills, 51*, 343–354.

Schachter, S. (1971). *Emotion, obesity, and crime*. New York: Academic Press.

Schaie, K. W. (1965). A general model for the study of developmental problems. *Psychological Bulletin, 64*, 92–107.

Schouten, J. W., & McAlexander, J. H. (1995). Subcultures of consumption: An ethnography of the new bikers. *Journal of Consumer Research, 22*, 43–61.

Schuler, H. (1982). *Ethical problems in psychological research*. New York: Academic Press.

Seligman, M. E. P. (1970). On the generality of the laws of learning. *Psychological Review, 77*, 406–418.

Seligman, M. E. P., & Hager, J. L. (1972). *Biological boundaries of learning*. New York: Appleton-Century-Crofts.

Shaffer, D. (1985). *Developmental psychology: Theory, research, and applications*. Monterey, CA: Brooks/Cole.

Shanks, N. (2003). Animal rights in the light of animal cognition. *Social Alternatives, 22*, 12–18.

Shapiro, F. (1989). Eye movement desensitization: A new treatment for post-traumatic stress disorder. *Journal of Behavioral Therapy and Experimental Psychiatry, 20*, 211–217.

Sheridan, C. E. (1979). *Methods of experimental psychology*. New York: Holt, Rinehart & Winston.

Shin, Y. H., (1999). The effects of a walking exercise program on physical function and emotional state of elderly Korean women. *Public Health Nursing, 16*, 146–154.

Shohat, M., & Musch, J. (2003). Online auctions as a research tool: A field experiment on ethnic discrimination. *Swiss Journal of Psychology, 62*, 139–145.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–428.

Sidman, M. (1953). Two temporal parameters of the maintenance of avoidance behavior by the white rat. *Journal of Comparative and Physiological Psychology, 46*, 253–261.

Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology*. New York: Basic Books.

Sieber, J. E., Iannuzzo, R., & Rodriguez, B. (1995). Deception methods in psychology: Have they changed in 23 years? *Ethics and Behavior, 5*, 67–85.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York: McGraw-Hill.

Sigelman, L. (1981). Question order effects on presidential popularity. *Public Opinion Quarterly, 45*, 199–207.

Signal, T. D., & Taylor, N. (2006). Attitudes to animals in the animal protection community compared to a normative community sample. *Society and Animals, 14*, 265–275. doi: 10.1163/156853006778149181

Silverman, I., Shulman, A. D., & Weisenthal, D. L. (1970). Effects of deceiving and debriefing psychological subjects on performance in later experiments. *Journal of Personality and Social Psychology, 14*, 203–212.

Simpson, S. S., Bouffard, L. A., Garner, J., & Hickman, L. (2006). The impact of legal reform on the probability of arrest in domestic violence cases. *Justice Quarterly, 23*, 297–316.

Singer, P. (1975). *Animal liberation: A new ethics for our treatment of animals*. New York: Avon Books.

Singer, P. (2002). *Animal liberation*. New York: HarperCollins Publishers.

Skinner, B. F. (1949). Are theories of learning necessary? *Psychological Review, 57*, 193–216.

Slovic, P., & Fischoff, B. (1977). On the psychology of experimental surprise. *Journal of Experimental Psychology: Human Perception and Performance, 3*, 544–551.

Smit, H. J., & Rogers, P. J. (2000). Effects of low doses of caffeine on cognitive performance, mood and thirst in low and higher caffeine consumers. *Psychopharmacology, 152*, 167–173.

Smith, S. L., Lachlan, K., & Tamborini, R. (2003). Popular video games: Quantifying the presentation of violence and its context. *Journal of Broadcast and Electronic Media, 47*, 58–76.

Smith, S. S., & Richardson, D. (1983). Amelioration of deception and harm in psychological research: The important role of debriefing. *Journal of Personality and Social Psychology*, *44*, 1075–1082.

Smith, T. E., Sells, S. P., & Clevenger, T. (1994). Ethnographic content analysis of couple and therapist perceptions in a reflecting team setting. *Journal of Marital and Family Therapy*, *20*, 267–286.

Snowdon, C. T. (1983). Ethnology, comparative psychology, and animal behavior. *Annual Review of Psychology*, *34*, 63–94.

Solomon, S., Greenberg, J., & Pyszczynski, T. (1991). A terror management theory of social behavior: The psychological functions of self-esteem and cultural worldviews. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 24, pp. 93–159). New York: Academic Press.

Stanovich, K. E. (1986). *How to think straight about psychology*. Glenview, IL: Scott, Foresman.

Steinberg, J. A. (2002). Misconduct of others: Prevention techniques for researchers. *American Psychological Society Observer*. Retrieved from http://www.psychologicalscience.org/observer/0102/misconduct.html.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677–680. Stevenson, M. R. (1995). Is this the definitive sex survey? *Journal of Sex Research*, *32*, 77–91.

Stevenson, M. R. (1995). Is this the definitive sex survey? *Journal of Sex Research*, *32*, 77–91.

Steward, K. K., Carr, J. E., Brandt, C. W., & McHenry, M. M. (2007). An evaluation of the conservative dual-criterion method for teaching university students to visually inspect AB-design graphs. *Journal of Applied Behavior Analysis*, *40*, 713–718.

Stolle, D. P., Robbennolt, J. K., Patry, M., & Penrod, S. D. (2002). Fractional factorial designs for legal psychology. *Behavioral Sciences and the Law*, *20*, 5–17.

Strayer, D. L., & Drews, F. A. (2007). Cell phone induced distraction. *Current Directions in Psychological Science*, *16*, 128–131.

Streiner, D. L. (2006). Building a better model: An introduction to structural equation modeling. *Canadian Journal of Psychiatry*, *51*, 317–324.

Strunk, W., & White, E. B. (1979). *The elements of style* (3rd ed.). New York: Macmillan.

Suls, J., & Martin, R. (2009). The air we breathe: A critical look at practices and alternatives in the peer-review process. *Perspectives on Psychological Sciences*, *4*, 40–50.

Swami, V., Furnham, A., & Christopher, A. N. (2008). Free the animals? Investigating attitudes toward animal testing in Britain and the United States. *Scandinavian Journal of Psychology*, *49*, 249–276. doi: 10.1111/j.1467-9450.2008.00636.x

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Boston: Allyn and Bacon.

Tabachnick, B. G., & Fidell, L. S. (2006). *Using multivariate statistics* (5th ed.). Boston: Allyn and Bacon.

Tanford, S. L. (1984). *Decision making processes in joined criminal trials*. Unpublished doctoral dissertation, University of Wisconsin, Madison.

Tanner, W. P., Jr., Swets, J. A., & Green, D. M. (1956). *Some general properties of the hearing mechanism* (Tech. Rep. No. 30). Ann Arbor: University of Michigan, Electronic Defense Group.

Tatsuoka, M. M. (1971). *Multivariate analysis: Techniques for educational and psychological research*. New York: Wiley.

Taylor, S. E. (1989). *Positive illusions: Creative self-deception and the healthy mind*. New York: Basic Books.

Therrien, K., Wilder, D. A., Rodriguez, M., & Wine, B. (2005). Preintervention analysis and improvement of customer greeting in a restaurant. *Journal of Applied Behavior Analysis*, *38*, 411–415.

Thorndike, R. M. (1978). *Correlational procedures for research*. New York: Gardner Press.

Tinbergen, N. (1951). *The study of instinct*. Oxford: Clarendon Press.

Treadway, M., & McCloskey, M. (1987). Cite unseen: Distortions of Allport and Postman's rumor study in the eyewitness testimony literature. *Law and Human Behavior*, *11*, 19–26.

Trujillo, N. (1993). Interpreting November 22: A critical ethnography of an assassination site. *Quarterly Journal of Speech*, *79*, 447–466.

Tsang, J-A. (2006). Gratitude and prosocial behaviour: An experimental test of gratitude. *Cognition and Emotion*, *20*, 138–148.

Tucker, W. H. (1997). Re-reconsidering Burt: Beyond a reasonable doubt. *Journal of the History of the Behavioral Sciences*, *33*, 145–162.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

U.S. Department of Commerce (2008). Networked nation: Broadband in America, 2007. Retrieved from http://www.ntia.doc.gov/reports/2008/NetworkedNationBroadbandinAmerica2007.pdf.

U.S. Department of Health and Human Services (2005). Basic HHS policy for protection of human research subjects. Retrieved from http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.htm.

U.S. Public Health Service. (2002). Public Health Service policy on humane care and use of laboratory animals. Retrieved from http://grants1.nih.gov/grants/olaw/olaw.htm.

Ullman, D., & Jackson, T. (1982). Researchers' ethical concerns: Debriefing from 1960–1980. *American Psychologist*, *37*, 972–973.

Underwood, M. K., Scott, B. L., Galperin, M. B., Bjornstad, G. J., & Sexton, A. M. (2004). An observational study of social exclusion under varied conditions: Gender and developmental differences. *Child Development*, *75*, 1538–1555.

Unger, R. K. (1983). Through the looking glass: No wonderland yet. *Psychology of Women Quarterly*, 8, 9–32.

Unger, R., & Crawford, M. (1992). *Women and gender: A feminist psychology*. New York: McGraw-Hill.

Vandell, D. L., & Hembree, S. E. (1994). Peer social status and friendship: Independent contributors to children's social and academic adjustment. *Merrill-Palmer Quarterly*, 40, 461–477.

Velleman, P. F., & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *American Statistician*, *47*, 65–72.

Vinacke, W. E. (1954). Deceiving experimental subjects. *American Psychologist*, 9, 155.

Vogel, D., & Wester, S. (2003). To seek help, or not to seek help: The risks of self-disclosure. *Journal of Counseling Psychology*, 50, 351–361.

Vollmer, T. A., & Bourret, J. (2000). An application of the matching law to evaluate the allocation of two- and three-point shots by college basketball players. *Journal of Applied Behavior Analysis*, *33*, 137–150.

Wadsworth, B. J. (1971). *Piaget's theory of cognitive development*. New York: McKay.

Walster, E., Berscheid, E., Abrahams, D. B., & Aronson, E. (1967). Effectiveness of debriefing after deception experiments. *Journal of Personality and Social Psychology*, 6, 371–380.

Walster, E., Walster, G. W., & Berscheid, E. (1978). *Equity theory and research*. Boston: Allyn and Bacon.

Warner, J. L., Berman, J. J., Weyant, J. M., & Ciarlo, J. A. (1983). Assessing mental health program effectiveness: A comparison of three client follow-up methods. *Evaluation Review*, *7*, 635–658.

Watson, J. B., & Rayner, R. (1920). Conditioned emotional reactions. Retrieved from http://psychclassics.yorku.ca/Watson/emotion.htm.

Weinfurt, K. P., & Bush, P. J. (1995). Peer assessment of early adolescents solicited to participate in drug trafficking: A longitudinal analysis. *Journal of Applied Social Psychology*, *25*, 2141–2157.

Westerlund, D., Granucci, E. A., Gamache, P., & Clark, H. B. (2006). Effects of peer mentors on work-related performance of adolescents with behavioral and/or learning disabilities. *Journal of Positive Behavior Interventions*, 8, 244–251.

Williams, C. D. (1959). The elimination of tantrum behavior by extinction procedures. *Journal of Abnormal and Social Psychology*, 59, 269.

Wilson, D. W., & Donnerstein, E. (1977). Guilty or not guilty? A look at the simulated jury paradigm. *Journal of Applied Social Psychology*, 7, 175–190.

Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.

Winkel, G. H., & Sasanoff, R. (1970). An approach to objective analysis of behavior in architectural space. In H. M. Proshansky, W. H. Ittelson, & L. G. Rivlin (Eds.), *Environmental psychology: Man and his environment* (pp. 619–630). New York: Holt, Rinehart & Winston.

Wolchik, S. A., Spencer. S. L., & Lisi, I. S. (1983). Volunteer bias in research employing vaginal measures of sexual arousal. *Archives of Sexual Behavior*, *12*, 399–408.

Wong, D., and Baker, C. (1988). Pain in children: Comparison of assessment scales. *Pediatric Nursing*, *14*, 9–17.

Wood, C. (1979). The I-knew-it-all-along effect. *Journal of Experimental Psychology: Human Perception and Performance*, *43*, 345– 353.

World Medical Association Declaration of Helsinki (1964). Retrieved from http://www.wma.net/e/policy/b3.htm.

Wozniak, R. H. (1999). Oskar Pfungst: *Clever Hans (The horse of Mr. von Osten)* (1907; English 1911). Retrieved from http://www.thoemmes.com/psych/pfungst.htm.

Wright, D. E., Titus, S. L., & Cornelison, J. B. (2008). Mentoring and research: An analysis of research mentoring in closed ORI cases [Electronic version]. *Science, Engineering and Ethics*, *14*, 323–336. doi: 10.1007/s11948-008-9074-5

Wuensch, K. L., & Poteat, G. M. (1998). Evaluating the morality of animal research: Effects of ethical ideology, gender, and purpose. *Journal of Social Behavior and Personality*, *13*, 139–150.

Wundt, W. (1897). *Outlines of psychology*. Retrieved from http://psychclassics.yorku.ca/Wundt/Outlines/

Yaremko, R. M., Harari, H., Harrison, R. C., & Lynn, E. (1982). *Reference handbook of research and statistical methods*. New York: Harper & Row.

Zeidner, M. (2006). Gender group differences in coping with chronic terrorism: The Israeli scene. *Sex Roles*, *54*, 297–310.

CREDITS

Page 41 Fig. 2-1 reprinted with permission from K. J. Lorenz, "The Comparative Method in Studying Innate Behavior Patterns," Symposia of the Society for Experimental Biology, 4; adapted from Fig. 2.4 in D. A. Dewsbury, *Comparative Animal Behavior.* Copyright © 1950, 1978 by Society for Experimental Biology and McGraw-Hill Company respectively.

Page 75 Fig. 3-3 reprinted with permission from Midwestern Psychological Association.

Page 107 Fig 4-1 reprinted with permission from Animal Behaviour, Vol. 77, Issue 4, Cartel ten Cate, Niko Tinbergen and the red patch on the herring's beak, Page 10, Copyright © 2009, with permission from Elsevier.

Page 143 Fig 5-2 reprinted from Pediatric Nursing, 1988, Volume 14, Number 1, pp. 9–17. Reprinted with permission of the publisher, Jannetti Publications, Inc., East Holly Avenue, Box 56, Pitman, NJ 08071-0056; (856) 256-2300; FAX (856) 589–7463; Web site: www.pediatricnursing.net; For a sample copy of the journal, please contact the publisher.

Page 189 Fig. 6-4 reprinted from C. D. Williams, "The Elimination of Tantrum Behavior by Extinction Procedures," *Journal of Abnormal and Social Psychology*, *59*, 269. Copyright © 1959 by the American Psychological Association. Reprinted with permission.

Page 297 Fig. 10-2 reprinted with permission from P. E. Gold, "Sweet Memories," *American Scientist*, p. 153. Copyright © 1987 by Sigma Xi, the Scientific Society, Inc. Reprinted by permission.

Page 317 Fig. 10-6 reprinted from L. P. Peterson and M. J. Peterson, "Short-term Retention of Individual Verbal Items," *Journal of Experimental Psychology*, *58*, 193–198. Copyright © 1959 by the American Psychological Association. Reprinted with permission.

Page 366 Fig 12-2 reprinted with permission from Melissa M. Anglesea, Hannah Hoch, and Bridget A. Taylor, "Reducing Rapid Eating In Teenagers With Autism: Use of a Pager Prompt" from *Journal of Applied Behavior Analysis*, *41*, 107–111 c 2008. Copyright © 2008.

Page 368 Fig 12-3 reprinted with permission from From R. V. Hall, D. Lund, and J. Jackson, "Effects of Teacher Attention on Study Behavior" from *Journal of Applied Behavior Analysis*, *1*, 1–12, c 1968. Copyright © 1968.

Page 374 Fig 12-5 K. Therrien, D. A. Wilder, M. Rodriguez, and B. Wine (2005), "Preintervention Analysis and Improvement of Customer Greeting in a Restaurant," *Journal of Applied Behavior Analysis*, 38, 411–415. Copyright © 2005.

Page 375 Fig 12-6 reprinted with permission from B. Abbott and P. Badia, "Choice for Signaled Over Unsignaled Shock as a Function of Signal Length," *Journal of the Experimental Analysis of Behavior*, *32*, 409–417. Copyright © 1979 by the Society for the Experimental Analysis of Behavior.

Page 379 Fig. 12-9 from "Effects of Peer Mentors on Work-related Performance of Adolescents with Behavioral and/or Learning Disabilities" by D. Westerlund, E. A. Granucci, P. Gamache, and H. B. Clark (2006). *Journal of Positive Behavior Interventions* 8, 244–251. Copyright © 2006 by PROED. Reprinted with permission.

Page 381 Fig 12-10 reprinted with permission from W. Palya, D. Walter, R. Kessel, R. Lucke, "Investigating Behavioral Dynamics with a Fixed-Time Extinction Schedule and Linear Analysis." *Journal of the Experimental Analysis of Behavior* (1996), 66, 391–409. Copyright (1996) by the Society for the Experimental Analysis of Behavior, Inc.

Page 451 Fig. 14-6 from D. F. Chang and S. Sue, "The Effects of Race and Problem Type on Teachers' Assessments of Student Behavior," *Journal of Consulting and Clinical Psychology, 71*, 235–242. Copyright © 2003 by the American Psychological Association. Reprinted with permission.

Page 481 Table 15-2 reprinted with permission from D. Vogel and S. Wester, "To Seek Help, or Not to Seek Help: The Risks of Self-disclosure," *Journal of Counseling Psychology, 50*, 351–361. Copyright © 2003 by the American Psychological Association. Reprinted with permission.

Pages A-7–A-11 Table 2, Table 3A, Table 3B reprinted with permission from E. S. Pearson and H. O. Hartley, *Biometrika: Tables for Statisticians*, Tables 12 & 18, 1966, Volume 1, 3rd edition, by permission of Oxford University Press.

Pages A-12–A-13 Tables 4A and 4B reprinted with permission from R. C. Milton, "An Extended Table of Critical Values for the Mann-Whitney (Wilcoxon) Two-sample Statistic," *Journal of the American Statistical Association, 59*, 925–934. Copyright © 1964 by the American Statistical Association.

Page A-14 Table 5 reprinted with permission from Bruning/Kintz, Computational Handbook of Statistics, Table "Areas under the Normal Curve", © 1997 HarperCollins College Publishers. Reproduced by permission of Pearson Education, Inc.

Page A-15 Table 6 reprinted with permission from E. S. Pearson & H. O. Hartley, *Biometrika: Tables for Statisticians*, Table 8, 1966, Volume 1, 3rd edition, by permission of Oxford University Press.

Page A-16 Table 7 reprinted with permission from A. L. Edwards, *Experimental Design In Psychological Research* (Fifth Edition). Copyright © 1985 David L. Edwards. Reprinted by permission of the author.

NAME INDEX

Abbott, B. B., 141, 191, 192, 362, 373, 374, 375, 387
Abrahams, D. B., 182
Adair, J. G., 150–151
Agresti, A., 490
Aguinis, H., 179–180
Ajzen, I., 491
Allport, G. W., 69, 258
Anastasi, A., 350
Anderson, C. A., 106
Anderson, J. L. R., 129
Anderson, N., 38–39
Anglesea, M. M., 365, 366
Anolli, L., 166
Applebaum, M. I., 352, 353
Aragon, A. S., 185
Arellano-Galdamas, F. J., 178
Aristotle, 47
Aronson, E., 122, 135, 182, 184
Asher, H. B., 466, 472, 494, 496
Ax, A. F., 199

Badia, P., 191, 192, 225, 361, 362, 363, 364, 373, 374, 375
Bakeman, R., 224–225, 226, 230, 231–232
Baker, C., 143
Balcetis, E., 298–299
Bandilla, W., 270
Bankert, E., 201
Banks, C., 121, 181
Bansal, S., 173
Banse, R., 123–124
Barnett, N. J., 166
Barrett, J., 215
Baumans, V., 189–190
Baumeister, R. F., 60–61
Baumrind, D., 180
Begin, A., 173
Bell, R., 216, 217, 218
Bem, D. J., 62
Bender, K., 271
Bennett, S., 476
Berenson, C., 270
Berg, B. L., 238, 239, 240, 241, 243
Berman, J. J., 270, 273
Berman, M. G., 322
Bernstein, D. M., 344, 345
Berscheid, E., 32–33, 128, 182
Bethell, C., 167
Bjornstad, G., 223–224, 225, 226, 227
Blatchley, B., 179, 180
Block, G., 193
Bolstein, R., 270
Bolt, M., 192
Boney-McCoy, S., 166
Bordens, K. S., 58, 59, 139, 140, 151, 395, 397, 401, 402
Borszcz, G., 148, 157
Bosnjak, M., 270
Bouffard, L., 340
Bourret, J., 34
Bowden, M., 271
Bowers, D., 476
Boynton, P., 173
Braithwaite, R. B., 22
Brandt, C. W., 365
Bray, J. H., 468, 496
Brewer, M. B., 286
Broad, W., 215, 216, 217, 218
Broca, P. P., 244–245
Brody, J. L., 185
Brown, R., 14–15
Brown, S. R., 146
Bruck, M., 206
Bruckman, A., 207, 208
Bruening, S., 219
Bruning, J. L., A14
Buehler, R., 183
Bulte, I., 386
Burt, C., 219–220
Bush, P. J., 243
Bush-Armendariz, N. B., 271
Butler, B. E., 220

Cameron, L., 331–332
Campbell, D. T., 114, 116, 118, 182, 339, 340, 341, 342, 344, 345, 346, 347
Carlsmith, J. M., 122, 184
Carnahan, T., 176
Carr, J. E., 365
Carroll, R. T., 9
Carson, H., 17–18
Castellan, N. J., 453, 455, 456
Ceci, S. J., 95, 206
Chabris, C. F., 246
Chang, D. F., 449–450, 451
Charvat, W., 537
Chassan, J. B., 386
Cheam, A., 183
Cheney, D., 318
Chomsky, N., 39
Christopher, A. N., 193
Church, A. H., 270
Cialdini, R. B., 7–8, 23, 24
Ciarlo, J. A., 270, 273
Clark, H., 378, 379
Clevenger, T., 238
Cogoni, D., 14
Cohen, J., 458
Cohen, J. M., 201, 202
Cohen, R. J., 133, 275
Coker, R., 10, 11
Conrad, E., 98, 227
Cooper, H. M., 255
Cooper, J. A., 201
Cooperman, E., 171
Cornelison, J., 218
Correll, J., 147
Couper, M. P., 270

Crano, W. D., 286
Crawford, M., 97
Crews, F., 534, 537
Crick, F., 218
Critchfield, T. S., 34
Culbertson, S., 361, 362, 363, 364

Dalal, R., 166, 167
Darwin, C., 358
Davidson, B., 237–238
Davidson, P. R., 12
Davis, A. J., 247
Davis, M. H., 176
DeBeuckelar, A., 270
DeSantis, A. D., 241–242
Descartes, R., 21–22, 190
DeWall, N. C., 60–61
Dewsbury, D. A., 41
Dill, K. E., 106
Dillman, D. A., 263, 264, 267, 268, 269, 270, 281
DiNitto, D. M., 271
Donnerstein, E., 122
Drews, F. A., 25, 26, 28, 111–112
Dunning, D., 298–299
Dyer, J., 271

Ebbinghaus, H. E., 53, 357, 358
Edwards, A. L., 266, 310, 324, A16
Ellis, C., 238
Engle-Friedman, M., 173
Epley, N., 182–183

Fancher, R. E., 18, 45, 133, 358
Fanselow, M. S., 167–168
Fechner, G. T., 357
Feild, H. S., 166
Feldman, M. S., 140
Festinger, L., 42, 43, 62
Fidell, L. S., 461, 466, 469, 470, 471, 472, 473, 474, 475, 476, 479, 480, 482, 483, 484, 487, 488, 489, 490, 496
Finlay, B., 490
Fiorillo, J., 167
Fireman, B., 182
Fischer, H., 129
Fischoff, B., 192
Fishbein, M., 491
Fisher, R., 359
Fisher, W. W., 365
Fiske, D. W., 95
Fiske, S. T., 211
Fleiss, J. L., 233
Fleming, C., 271
Flourens, P., 45
Fogg, L., 95
Folkman, J. R., 165
Forscher, B. K., 52–53
Foy, E., 15
Fredrikson, M., 129
Freedman, J. L., 181
Friedman, P., 218
Fry, D. P., 238
Furmark, T., 129
Furnham, A., 193

Gabriel, U., 123–124
Gagnon, J. H., 273
Gaither, G. A., 174
Galen, M. A, 189
Gall, F., 45, 52
Galperin, M. B., 223–224, 225–226, 227
Galton, F., 358
Gamache, P., 378, 379
Gamson, W. A., 182
Garcia, J., 91–92
Garner, J., 340
Garson, D. G., 495
Geggie, D., 217, 218
Gerbing, G. W., 491
Gibbon, J., 42–43
Gilbert, D. T., 318, 319, 320, 321
Glass, G. V., 254
Glickman, M. E., 246
Gluck, J. P., 185
Gold, P. E., 296, 297, 298
Goldiamond, I., 145
Goldstein, I., 7
Goldstein, J. H., 151
Goodall, J., 237
Goodman, B., 1–2, 5, 7, 23, 105
Goodstadt, B. E., 151
Gordon, B. N., 206
Gottman, J. M., 224–225, 226, 230, 231–232
Granucci, E., 378, 379
Gravetter, F. J., 391, 421, 422, 424, 431, 438, 442, 443, 447, 453, 457, 458
Green, D. M., 383, 384, 385
Green, R., 128
Greenberg, B. S., 247, 248, 249
Greenberg, J., 60
Greene, E., 151
Greenwald, A. G., 147
Greenwood, K. M, 365
Gretz, H. F., III, 7
Grice, G. R., 312
Guzinski, G. M., 7

Hager, J. L., 92
Hagger, M., 169
Hall, D., 533, 534, 537
Hall, J. A., 176
Hall, R. V., 367–368
Hamby, S., 166
Haney, C., 121, 181
Harari, H., 4, 23, 453, 455
Harari, O., 453, 455
Harrison, R. C., 4, 23
Harsh, J., 362
Hartley, H. O., A7, A8, A10, A15
Hartley, L. R., 336–337, 338
Helmstetter, F. J., 167–168
Hembree, S. E., 243
Hempel, C. G., 46
Hendryx, M., 167
Henle, C. A., 179–180
Herrmann, D., 206
Herrnstein, R. J., 34
Hershberger, S. L., 495
Hertwig, R., 178
Hess, D., 440, 441
Hickman, L., 340
Hickson, L., 237, 238
Higbee, K. L., 165
Hite, S., 278–279
Hoch, H., 365, 366
Holmes, D. S., 180, 184
Holsti, O. R., 247–248
Hooke, R., 278
Horowitz, I. A., 59, 140, 176, 182, 247, 395, 397, 401, 402
Huang, H-M., 166
Hubbard, M., 183
Huck, S. W., 114, 326
Hudson, J. M., 207–208
Huff, C., 182–183
Hull, C., 51
Hunter, J. E., 250, 472, 491, 496

Iannuzzo, R., 178

Jackson, D., 367–368
Jackson, T., 182
James, J. M., 270
Janis, I., 181
Jay, P., 215
Jones, R. A., 207
Jonides, J., 322
Joynson, R. B., 220
Judd, C. M., 147

Kaczmirek, L., 270
Kaiser, M. K., 488
Kalichman, M., 218
Kalton, G., 261, 267, 286, 287
Kanuk, L., 270
Kaplan, S., 173, 322
Kardes, F., 166
Kassam, K. S., 318, 319–321
Katz, J., 201, 202, 205
Kaukiainen, A., 243–244
Kazdin, A. E., 365, 386
Kelley, M. E., 365
Kelly, G., 5
Kelman, H. C., 179
Keppel, G., 310, 334, 444, 445, 446, 447, 449, 451, 452, 458, 460, 488
Kessel, R., 380
Key, W. B., 113
Kintz, B. L., A14
Kish, L., 281, 283–285, 287
Kjerulff, K. H., 7
Knickman, J., 167
Koelling, R. A., 91–92
Kolata, G., 273
Krantz, J. H., 166, 167
Kreutzer, J., 440, 441
Krippendorff, K., 247
Kruse, C. R., 193
Kuczynski, E., 7
Kuhn, T. S., 49
Kunda, Z., 37

Labruna, V., 173
Lachlan, K., 249, 250
LaFrance, M., 73
Lagerspetz, K., 243–244
Landon, A., 278
Landy, E., 135
Laney, C., 344, 345
Langenberg, P. W., 7
Lansky, D., 167
Latané, B., 44
Laumann, E. O., 273
Laverty, D., 1
Leaton, R. N., 148, 157
Lee, R. E., 9
Leggett, G., 537
Leikas, S., 173
Lepper, M. R., 183
Levine, M. S., 466, 484
Lewis, J. E., 193
Libby, L., 318, 319, 320
Liepman, M. R., 173
Lievens, F., 270
Lilienfeld, S. O., 9, 10
Lindsey, D., 95
Link, M. W., 167
Lisi, I. S., 176
Loftus, E. F., 5, 151, 206, 344, 345
Lohr, J. M., 9
Lomas, J. E., 365
Longino, H., 96
Lönnqvist, J., 173
Lönnqvist, J-E., 173
Lord, F. M., 138
Lorenz, K., 40, 41, 42
Lothert, J., 176
Lucke, R., 380, 381
Lund, D., 367–368
Lynn, E., 4, 23
Lynn, S. J., 9

Macaulay, D., 17–18
Mahoney, M. J., 94–95
Mandel, F. S., 173
Mann, L., 181
Manolov, R., 386
Mans, G., 174
Marcoulides, G. A., 495
Marcus, B., 171–172
Martin, E., 32, 94
Martins, B. K., 34
Marwitz, J., 440, 441
Matfield, M., 190
Matyas, T. A., 365
Maul, T., 227
Maxwell, S. E., 468, 496
Mayo, C., 73
McAlexander, J. H., 238
McCabe, G. P., 409
McCain, J., 258–259
McCall, R. B., 352, 353
McCartt, A. T., 105
McCloskey, M., 69
McEnvoy, S. P., 105
McFarland, C., 183
McFarland, S., 176, 267–268
McGhee, D. E., 147
McGraw, K. O., 233
McHenry, M. M., 365
McNemar, Q., 165
Mead, C. D., 537
Meier, B. P., 174
Mengele, J., 200, 201
Meyer, M., 176
Michael, R. T., 273
Milgram, S., 149, 180
Millard, R. J., 165
Miller, J., 14
Miltenberger, R. G., 153–154
Milton, R. C., A12, A13
Mitchell, K. V., 176
Mitchell, S. K., 230
Mokdad, A., 167
Montee, B. B., 153–154
Mook, D. G., 118, 168
Moore, D. S., 409
Morris, E. K., 344, 345
Moser, C. A., 261, 267, 286–287
Mosteller, F., 138
Musch, J., 170
Myers, D. G., 97, 192

Neisser, U., 139
Nerb, J., 36–37
Neubarth, W., 270
Niedźwieńska, A., 183
Nirenberg, T. D., 173
Nowlis, V., 128
Nunnally, J. C., 265

Obama, B., 258–259
O'Brien, K. R., 179, 180
O'Brien, R. G., 488
Ockham, W., 47
Oczak, M., 183
Oei, A., 336–337, 338
Ogloff, J. R. P., 5
Onghena, P., 386
Orbell, S., 169
Orne, M. T., 151
Ornstein, P. A., 206
Ortmann, A., 178

Pagano, R. R., 391, 422, 446
Palya, W., 380, 381, 382
Park, B., 147
Parker, K. C. H., 12
Parramore, M. M., 495
Party, M., 324
Pasupathi, M., 25, 26
Paunonen, S., 173
Pavlov, I., 92, 120, 358
Pearson, E. S., A7, A8, A10, A15
Pearson, K., 358
Pelcovitz, D., 173
Penrod, S. D., 324
Peplau, L. A., 98
Peters, D. P., 95
Peterson, L. R., 316–317
Peterson, M. J., 316–317
Petrulis, J., 220
Pfungst, O., 153
Piaget, J., 43–44, 59, 71, 235

Pittenger, D. J., 178, 208–209
Platt, J. R., 49–50
Plous, S., 185, 193
Poisson, R., 215
Postman, L., 69
Poteat, G. M., 186
Powers, W. T., 381, 382
Prelec, D., 34
Probst, T., 346–347
Pyszczynski, T., 60

Rayner, R., 198, 199, 200
Redelmeier, D. A., 105
Reed, D. D., 34
Renfrey, G., 12
Rescorla, R. A., 35
Resnick, J. H., 181
Reynolds, G. S., 368, 369
Rhodes, J. C., 7
Richardson, D., 182, 183
Riva, G., 166
Robbennolt, J. K., 324
Roberts, J. V., 254
Rodriguez, B., 178
Rodriguez, M., 373, 374
Rogers, P. J., 303–305, 308
Rogers, T. B., 274–276
Roosevelt, F. D., 278
Roscoe, J. T., 453, 455, 479
Rosenthal, R., 91, 154, 156, 171, 172, 173–174, 175, 177, 251, 252, 253, 254, 255
Rosnow, M. 538
Rosnow, R. L., 151, 171, 172, 173–174, 175, 177, 538
Ross, L., 183
Rothschild, B. H., 182
Roy, A. K., 226
Runyon, R. P., 225
Rutland, A., 331–332
Rytina, S., 182

Sagar, H. A., 58–59
Saini, J., 7
Saks, M., 93
Salmivalli, C. A., 243–244
Sandler, H. M., 114, 326
Sasanoff, R., 121
Saunders, D. R., 171
Schaie, K. W., 353
Schmidt, B. H., 250
Schofield, J. W., 58–59
Schouten, J. W., 238
Schuler, H., 178–179, 185, 200, 203
Schütz, A., 171–172
Schwartz, T., 147, 181
Scott, B. L., 223–224, 225–226, 227
Seligman, M. E. P., 20, 92, 191
Sellbom, M., 174
Sells, S. P., 238
Sexton, A. M., 223–224, 225–226, 227
Shaffer, D., 350
Shanks, N., 190
Shaprio, F., 11–12
Sheridan, C. E., 146
Shin, Y. H., 342, 343
Shohat, M., 170
Shrout, P. E., 233
Shulman, A. D., 179
Sidman, M., 369, 376, 377, 387
Sieber, J. E., 178
Siegel, S., 453, 455, 456
Sigelman, L., 268
Signal, T. D., 193
Sills, E. S., 7
Silverman, I., 179
Simpson, S. S., 340
Singer, P., 190–192
Skinner, B. F., 45, 46, 53, 359
Skorton, D., 193
Skorton, S., 193
Slanas, A., 386
Slovic, P., 192
Smit, H. J., 303–305, 308
Smith, S. L., 249–250
Smith, S. S., 182, 183
Smith, T. E., 238
Snapp, T., 176
Snowdon, C. T., 186
Solomon, S., 60
Spada, H., 36–37
Spates, C. R., 12
Spence, K., 51
Spencer, S. L., 176
Sprague, R., 219
Stanley, J. C., 114, 116, 118, 339, 340, 341, 342, 344, 345, 346, 347, 352
Stanovich, K. E., 168
Steele, C., 81
Steinberg, J. A., 218–219
Stephenson, W., 146
Stevens, S. S., 135, 137, 138
Stevenson, M. R., 105, 277
Steward, K. K., 365
Stolle, D. P., 324
Strayer, D. L., 25, 26, 111–112
Stream, C., 174
Streiner, D. L., 495
Strunk, W., Jr., 537
Sue, S., 449–450, 451
Sugarman, D. B., 166
Suls, J. E., 94, 151
Swami, V., 193
Swencionis, J. K., 318, 319, 320, 321
Swerdlik, M. E., 133, 275
Swets, J. A., 383, 384, 385

Tabachnick, B. G., 461, 466, 469, 470, 471, 472, 473, 474, 475, 476, 479, 480, 482, 483, 484, 487, 488, 489, 490, 496
Tackett-Gibson, M., 271
Tamborini, R., 249–250
Tanford, S. L., 166
Tanner, W. P., Jr., 383, 384, 385
Tatsuoka, M. M., 474
Taylor, B. A., 365, 366
Taylor, N., 193
Taylor, S. E., 37
Teruzzi, T., 166
Therrien, K., 373, 374
Thorndike, E. L., 358
Thorndike, R. M., 477
Tibshirani, R. J., 105
Tinbergen, N., 57, 106–108
Titus, S., 218
Tolman, E., 51
Treadway, M., 69
Trujillo, N., 238
Tsang, J. A, 295–296
Tucker, W. H., 220
Tukey, J. W., 138, 408
Tuulio-Henrikkson, A., 173

Ullman, D., 182
Underwood, M. K., 223–224, 225–226, 227
Unger, R. K., 97, 98

Vandell, D. L., 243
Velleman, P. F., 138
Verkaslo, M., 173
Vidmar, N., 5
Vinacke, W. E., 200
Vogel, D., 481
Volmer, T. A., 34
Von Osten, W., 153, 154

Wade, N., 215, 216, 217, 218
Wadsworth, B. J., 71

Wagner, A. R., 35
Wallnau, L. B., 391, 421, 422, 424, 431, 438, 442, 443, 447, 453, 457, 458
Walster, E., 32–33, 128, 182
Walster, G. W., 32–33, 128
Walter, D., 380, 381
Warner, J. L., 270, 273
Watson, J. B., 197–198, 200
Watson, J. D., 218
Weber, M., 357
Weiner, M., 173
Weinfurt, K. P., 243
Weisenthal, D. L., 179
Wester, S., 481
Westerlund, D., 378, 379
Weyant, J. M., 270, 273
White, E. B, 537
White, R. V., 453, 455
Wik, G., 129
Wilder, D. A., 373, 374
Wilkinson, L., 138
Williams, C. D., 188, 189
Wilson, D. W., 122
Wilson, T. D., 318, 319, 320, 321
Wincze, J. P., 173
Wine, B., 373, 374
Winer, B. J., 324, 446, 461
Winkel, G. H., 121
Wittenbrink, B., 147
Wittrock, D., 153–154
Wolchik, S. A., 176
Wong, D., 143
Wong, S. P., 233
Woo, H., 271
Wood, C., 192
Worall, L., 237–238
Wozniak, R. H., 153
Wright, D., 218
Wuensch, K. L., 186
Wundt, W., 40, 358

Yaremko, R. M., 4, 23
Yerkes, R., 18
Yoder, C., 206

Zeidner, M., 62–63
Zimbardo, P., 121, 181

SUBJECT INDEX

AB baseline design, 372
ABA baseline design, 360, 370, 372
ABAB baseline design, 360
Abstract, in APA style, 505–506
Academic Search Premier, 76
Accuracy of a measure, 132
Alpha level, 436, 444, 457
American Psychological Association (APA), 29, 71
  ethical guidelines (animals), 212–214
  ethical guidelines (humans), 170, 203, 204–205, 206, 209
  journals published by, 72
  research deception views of, 178–179
  writing style, 500–501
Analogical theories, 40–42
Analysis of covariance (ANCOVA), 452
Analysis of variance (ANOVA), 442–452
Anchors, 265
Animal Liberation Front (ALF), 193
Animal rights movement, 189–190
Animal subjects
  acquiring, 187–188
  alternatives to using, 194–195
  considerations for, 185–193
  cost-benefit assessment, 213–214
  ethical guidelines 212–214
  nonrandom sampling and, 167–168
  research integrity and, 214–220
APA Ethical Guidelines. *See* American Psychological Association
Apparatus subsection, APA style, 511
Applied research, 4, 62–63
Apprehensive attitude, 151
A *priori* comparisons, 444
Archival research, 245–246
Arithmetic average, 410
Asymptote curves, 402–403
Author note, APA style, 505
Authority, method of, 21
Automation, 157

Bar graphs, 400–401
Baseline designs
  baseline phase, 360
  behavioral baseline, 360
  characteristics of, 361
  choosing a stability criterion, 363–364
  dealing with uncontrolled variability, 366–368
  determining generality of findings, 368
  intersubject replication, 361
  intervention phase, 360
  intrasubject replication, 360
  judging differences in stable performance across phases, 365
  stability criterion, 360, 364–365
  transitional behavior and stability criterion, 364
Baseline designs, types of
  AB, 372
  ABA, 360, 370, 372
  ABAB, 360
  multifactor, 376
  multiple-baseline, 376–380
  single-factor, 372–376
Baseline phase, 360
Basic research, 4
Behavior, 6–8
Behavior baseline, 360
Behavioral categories, 224–225
Behavioral measures, 145
Behavioral sequences, 226
Belief-based explanations, 16–17
Belmont Report, 202–203
Beneficence, principle of, 203
Beta weights, 424, 480
Between-group variability, 442
Between-subjects designs
  and error variance, 291–294
  defined, 291
  matched-groups, 299–302
  randomized-multigroup, 298–299
  single-factor randomized, 294–299
Bias
  confirmation, 6
  experimenter, 152–156
  nonresponse, 269
  observational research and, 234–235
  sample, 278
  subject selection, 117
  volunteer, 171
Biased language, avoiding, 530–531
Bimodal distributions, 410
Bivariate linear regression, 422–424
Blind observer, 235
Blind techniques, 155
Boxplots, 416–417

Canonical correlation, 467, 483–484
Carryover effects, 305–314, 350
Case history, 244–245
Casual observations, 57, 58
Causal inference, 468–469
Causal relationships, 103–104, 492–493
Causes of behavior, exploring, 7–8
Ceiling effects, 142
Centroid, 482
Chi-square, 453–455

Children, as research subjects, 205–206
Circular explanations, 19
Classical conditioning, 34–35
Clever Hans phenomenon, 152–153
Closed-end items, 262–263
Cluster sampling, 284–285
Code sheet, example, 394
Coding schemes, 224
Coefficient alpha, 275
Coefficient of determination, 424–425
Coefficient of nondetermination, 425
Cohen's Kappa, 230–232
Cohort-sequential design, 353
Combined designs, 330–334
Commonsense explanations, 14–15
Communication, personal, 75
Comparisons, 444–445
Composite scale, 140
Computer modeling, 35–37
Concurrent validity, 134, 277
Confirmation bias, 6
Confirmational strategy, 48
Confirmatory factor analysis, 476
Confounding, 115, 325–327
Confusion matrix, 231
Construct validity, 134, 276
Content analysis, 246–249
Content footnote, APA style, 522
Content validity, 133, 276
Contingency tables
    chi-square for, 453
Contrast effects, 312
Control group, 109
Conventions, 74–75
Cooperative attitude, 151
Copy editor, 539–540
Copyright permission footnote, APA style, 522
Correlation matrix, 425
Correlational relationships, 103–104
Correlational research
  advantages, 106–108
  defined, 104
  directionality problem, 106
  experimental designs combined with, 468
  multivariate correlational designs, 467
  third-variable problem, 105–106
Cost-benefit assessment, 213–214
Counterbalancing, 308–310

Covariates, 335
Criterion variables, 105, 261, 467
Criterion-related validity, 134, 277
Critical region, 436–437
Critical value of a statistic, 436
Cross-sectional designs, 348–350
Cues, role attitude, 150–151
Curvilinear relationship, 419

Data
  collection, 228
  computer entry, 396–398
  data driven *versus* theory driven research, 51–53
  file, 398
  from multiple observers, 234
  generality of animal research, 188
  graphing, 400–406
  grouped *versus* individual, 399
  organizing, 392–399
  qualitative, 228
  quantitative, 228
  summary sheets, 392–396
  transformation, 460–463
Data summary sheets, 392–396
Data transformation, 460–461
Debriefing, 182–185, 209
Deception
  problems involved in using, 179–180
  solutions to, 181–185
  types of in research, 178–179
Declaration of Helsinki, 201
Deductive reasoning, 28
Degrees of freedom (*df*), 431
Dehoaxing, 201
Demand characteristics, 150–152
Demographics, 261
Demonstrations, experiments *versus*, 113–114
Dependent measures or variables
  adequacy, 141
  defined, 109
  multiple, 325, 468
  range effects, 142
  sensitivity, 141
  tailoring, 143–144
  types of, 144–147
Descriptive statistics
  defined, 391
  five-number summary, 416
  measures of association, 418–426
  measures of center, 410–413
  measures of spread, 413–416
Descriptive theories, 39–40
Designs. *See* Research designs

Developmental designs, 348–353
Differential carryover effects, 310
Digital object identifier (DOI), 519–520
Direct replications, 369
Directionality problem, 106
Disconfirmational strategy, 49
Discrete trials designs, 383–386
Discriminant analysis, 482–483
Discrimination learning, 144
Discussion section, APA style, 517
Dissertation Abstracts International, 76
DOI system number, 519
Domain of a theory, 43
Double-blind technique, 155
Drifting baselines, 370
Dummy coding, 392, 397
Duration method, 225
Dynamic designs, 380–382

EBSCOhost, 80, 83
Ecological validity, 139–140
Economic sample, 286
Editorial policy, 93
Editorial review process, 539–540
Effect size, 457–458
Empirical question, 64
Endogenous variables, 493
Equity theory, 32–33
Equivalent time series designs, 340–341
Error bars, 401
Error variance
  defined, 291
  in single-subject designs, 386
  sources of, 291–292
  ways of handling, 293–294
Ethical Principles of Psychologists and Code of Conduct (APA 2002), 203, 204–205, 206, 209
Ethical research practice
  animal subjects, 211–214
  evolution of, 200–206
  human participants, 197–211
  institutional policies and, 170
  internet and, 270–271
  reporting or publishing results, 542
  research reports and, 544
  science and, 214–220
  survey research and, 259–260
Ethnography, 238–242
Ethograms, 224
Event sampling, 227
Exogenous variables, 493

Expectancy effects, 154
Experimental designs
  between-subjects, 294–302
  confounding in, 325–327
  error variance in, 291–295
  factorial, 317–324
  multivariate, 325
  other group-based, 324–325
  single-subject (small *n*), 357–388
  types of, 290–291
  within-subjects, 303–317
Experimental error, 442
Experimental mortality, 117
Experimental realism, 122
Experimental research
  defined, 109
  characteristics of, 109–110
  strengths and limitations of, 112
  versus demonstrations, 113–114
Experimenter bias, 152–156
Explanatory relevance, 46
Exploded pie graph, 405
Exploratory data analysis (EDA), 391–392
Exploratory data collection (EDC), 102
Exploratory factor analysis, 476
External validity, 118–119, 165, 176–177, 361
Extraneous variables, 110–111

F ratio, 444, 448
Face validity, 133
Face-to-face interviews, 272–273
Factor analysis, 467, 474–476
Factor loading, 474
Factorial designs,
  higher-order, 323–324
  interactions, 320–321
  main effects, 320
  simple main effects, 321
  treatment order and, 312
  within-subjects, 322–323
Familywise errors, 445
Faulty inference, 17–19
Field research, 169–170
Field settings, 123–124, 169–170, 238–242
Field survey, 259–260
File drawer phenomenon, 91, 252
Finite population correction (fpc), 286–287
Five-number summary, 416
Floor effects, 142
Fractional factorial design, 324
Fraud in research, 214–220
Frequency distribution, 406–409
Frequency method, 225
Functional explanation, 37–38
Functional relationships, 402
Fundamental theories, 42–43

Gatekeepers and ethnography, 240
Generalization, 165
Generation effect, 350
Government regulations and ethics, 203, 205–206
Graphs
  APA style for, 524–525
  bar, 400–401
  boxplots, 416
  elements of, 400
  histograms, 407–408
  importance of, 405–406
  line, 402–403
  pie, 404
  scatter plots, 404
Graphics on questionnaires, 268
Group-administered surveys, 271–272
Grouped data, 399
Guide for the Care and Use of Laboratory Animals, 212–213

Habituation, 236–238
Habituation technique, 143
Heteroscedasticity, 471
Heuristic value, 44
Hierarchical regression, 479
Higher-order factorial designs, 323–324
Histograms, 407–408
History effect, 352
Homoscedasticity, 471
Human participants
  acquiring, 168–170
  APA guidelines, 170, 203, 204–205, 206, 209
  children as, 205–206
  considerations when choosing, 28, 162–168
  deception, 178–185
  field research, 169–170
  government regulations, 203, 205, 206
  informed consent, 198
  internet research and, 207–211
  reactivity in, 149–156
  research integrity and, 214–220
  tailoring measures to, 143–144
  voluntary participation and, 171–177
Hypotheses
  defined, 23, 54
  developing, 26–27, 98–99
  rival, 114
  testing, 102
  theory *versus*, 33–34
Hypothetico-deductive method, 51

Impact factor, 73
Implicit measures, 147
In vitro alternative to animal research, 194
Independent variable
  defined, 109
  quasi, 336–339
  treatment order as, 311–313
Indirect measures, 237
Individual sampling, 227
Inferential statistics
  alpha level, 436, 444
  alternatives to, 461–463
  data transformations and, 460–461
  defined, 430
  degrees of freedom, 431
  effect size, 457–458
  logic behind, 432–438
  nonparametric, 456
  one-tailed *versus* two-tailed tests, 436–438, 457
  parametric, 438–447
  parametric *versus* nonparametric, 431–432
  power, 457–459
  sample size, 457
  sampling distribution, 430–431
  sampling error, 431
  scale of measurement and, 138–139
  significance, 435–436
  single-subject designs and, 386–388
  statistical errors, 434–435
  statistical *versus* practical significance, 459
Information yielded by a measure, 138
Informed consent, 198, 207–208
Informed-consent form, 210
Ingenta*Connect*, 80
Institute for Scientific Information (ISI), 73
Institutional animal care and use committee (IACUC), 212–213
Institutional policies and ethics, 170

Institutional review board (IRB), 210–211
Instrumentation and internal validity, 117
Interaction, 320–321, 448–449
Internal validity, 114–118, 174–175, 360
Internet research
  deception in, 209
  ethical issues, 207–209
  ideas from, 75–76
  informed-consent issues, 207–209
  nonrandom sampling and, 166–167
  privacy, 208
  surveys, 270–271
Interquartile range, 415
Interrater reliability,
  Cohen's Kappa, 230–232
  defined, 229
  intraclass correlation, 233
  Pearson product-moment correlation, 232–233
  percent agreement, 230
Interrupted time series designs, 339–340
Intersubject replication, 361
Interval scale, 137
Intervals method, 226
Intervention phase, 360
Intraclass correlation (ICC), 233
Intrasubject replication, 360, 369
Introduction, APA style, 506–510
Irreversible changes, 310

Jewish Chronic Disease Hospital and ethics, 203, 205
Journal Citations Report, 73
Journals
  editorial policy, 93
  list of, 72–73
  refereed *versus* nonrefereed, 71–73
  review process, 90
  submissions to, 539–540
Journals in psychology, 72–73
JSTOR, 80, 81, 83,
Justice, principle of, 203

Laboratory setting, 120–122, 169
Latent variables, 495
Latin square ANOVA, 447
Latin square designs, 310
Law, 34
Law of parsimony, 47
Least-publishable-unit rule, 216
Least-squares regression line, 422
Library research, 77–82
Likert scale, 146, 266
Line graphs, 402–403
Linear regression,
  bivariate, 422–424
  least-squares regression line, 422
  prediction errors, 424
  regression weight, 422, 424
  residuals, 422
Linear transformations, 460
Linearity assumption, 469
Literature review, meta-analysis *versus* traditional, 250, 255
Literature, reviewing
  reasons for doing, 66–67
  sources for 67–76
Loglinear analysis, 489–490
Longitudinal designs, 350–353

Mahalanobis distance, 470
Mail surveys, 269–270
Main effects, 320, 448–449
Manipulation check, 158–159
Mann-Whitney *U* test, 455–456
Manuscript. *See* Research reports, preparing
Matched-groups designs, 299–302
Matched-multigroup design, 302
Matched-pairs design, 302
Materials subsection, APA style, 511
Maturation and internal validity, 116
Mean, 411–412
Measurement
  choosing scale of, 138–144
  error, 472–473
  Q-sort technique, 146
  scales of, 135–137
Measures
  accuracy, 132–133
  adequacy, 141–142
  behavioral, 145
  choosing, 138–139
  dependent, 109,
  ecological validity, 139–140
  established, 134–135
  implicit, 147
  physiological, 145–146
  range effects, 142
  reactive nature of, 149–156
  reliability, 130–132
  resistant, 409
  scale of measurement of, 135–137
  self-report, 146
  tailoring to participants, 143–144
  validity, 133–135
  variables and, 137–138
Measures of association, 232–233, 418–426
Measures of center, 410–413
Measures of spread, 413–416
Mechanistic explanations, 37–38
Median, 411
Meetings, 74–75
Meta-analysis, 249–255
Method of authority, 21
Method section, APA style, 510–513
Methods of inquiry, 21–25
Mixed designs, 330–332, 451–452
Mode, 419–411
Model, 34–37
Monotonic curves, 403
Multicollinearity, 472
Multifactor single-subject designs, 376
Multiple-author citations, APA style, 526–527
Multiple control group design, 298
Multiple *R*, 479–480
Multiple regression, 478–482
Multiple-baseline designs, 376–380
Multiple-observation effects, 352
Multiple *R*-square 479–480
Multistage sampling, 285
Multivariate analyses
  canonical correlation, 483–484
  discriminant analysis, 482–483
  factor analysis, 474–476
  multiple regression, 478–482
  multivariate analysis of variance (MANOVA), 468, 484–488, 496
  multiway frequency analysis, 489–490
  path analysis, 491–494
  structural equation modeling (SEM), 495
Multivariate analysis of variance (MANOVA), 468, 484–488, 496
Multivariate correlational techniques, 426, 467–469
Multivariate designs, 466–473
Multivariate strategy, 466
Multiway frequency analysis, 489
Mundane realism, 122

National Research Act, 202
National Research Council (1996), 212–213

Naturalistic observation, 236–238
Negative attitude, 151
Negatively accelerated curves, 402
Nested designs, 332–334
Nominal scales, 136
Nonequivalent control group design, 342–343
Nonexperimental research
  archival, 245–246
  bias in, 234–235
  case history, 244, 245
  content analysis, 246–249
  ethnography, 238–242
  observational, 223–236
  meta-analysis, 249–255
  naturalistic observation, 236–238
  research designs, 236–249
  sociometry, 243–244
Nonmonotonic curve, 403
Nonparametric design, 298
Nonparametric statistics, 432, 453–456
Nonparticipant observation, 238
Nonrandom sampling, 165–168
Nonrefereed journals, 71–73
Nonresponse bias, 269
Nonscience, 9
Normal distribution, 409
Normality assumption, 409, 438
  measures of center, 413
  measures of spread, 416
  Pearson *r*, 405
Nuremberg Code, 201

Oblique rotation, 475
Observational research
  behavior sequences, 226
  behavioral categories, 224–225
  biased sources, 234–235
  complexity, coping with, 226–229
  data collection, 235–236
  establishing reliability, 229–234
  multiple observers, 234
  quantifying behavior, 225–226
  single events, 226
Observations
  ethnography, 238–242
  naturalistic, 236–238
  in scientific method, 23
  systematic, 59–60
  unsystematic, 57–59
Observed value of a statistic, 436
Observer bias, 234–235

Occam's Razor, 47
Office of Research Integrity (ORI), 215
One-tailed test, 436–438
Open-ended items, 262
Operational definition, 64
Oral presentations, 540–541
Ordinal scale, 136
Ordinate (*x*-axis), 400
Orthogonal comparisons, 444
Orthogonal rotation, 475
Outliers, 409, 469–471

p-value, 444
Paper sessions (presentations), 74, 540–541
Parallel-forms reliability, 131
Parametric design, 298, 317, 372
Parametric statistics
  analysis of variance (ANOVA), 442–443
  assumptions underlying, 438
  defined, 431
  example, 440–441
  nonparametric *versus*, 431–432
  t test, 439–440
  with two samples, 438–439
  z test for proportions, 441
Parsimonious explanations, 13
Parsimony and theories, 47–48
Part correlation (semipartial correlation), 476, 477–478
Partial correlation, 476–477
Partial counterbalancing, 308–309
Partially open-ended items, 263
Participant observation, 239
Participants, *See* Human participants
Participants subsection, APA style, 510
Passive deception in research, 179
Path analysis, 491–494
Path coefficients, 494
Path diagram, 491
Pearson product-moment correlation (Pearson *r*), 232–233, 418–421
Peer review, 93–96
Percent agreement, 230
Per-comparison error, 445
Personal communications, 75
Phi coefficient (φ), 421
Photographs in manuscript, APA style, 524
Physiological measures, 145–146

Pie graphs, 404–405
Pilot study, 158
Plagiarism, 538
Planned comparisons, 444–445
Point-biserial correlation, 421
Population, 163
Positively accelerated curves, 402
Post hoc comparisons (unplanned comparisons), 445
Post hoc tests, 446
Poster sessions, 74–75, 541–542
Power of a statistical test, 457–459
Practical significance, 459
Predictive validity, 134, 277
Predictor variables, 105, 261, 467
Preference technique, 143
Pretest-posttest designs, 343–347
Primary sources, 69
Principal components analysis, 475–476
Principal factor analysis, 475–476
Prior consent to be deceived, 182
Probability pyramiding, 445
Procedure subsection, APA style, 511
Professional meetings, 74–75
Proportionate sampling, 282–283
Pseudoexplanations, 19–20
Pseudoscience, 9–10
PsycARTICLES, 77, 80, 81, 83
Psychological Abstracts, 78
Psychonomic Society, journals of, 73
PsycINFO, 77, 78–79, 82, 83
Publication Manual of the American Psychological Association (6th ed.), 501
Publication practices, 89

Q-sort methodology, 146
Qualitative approach 235
Qualitative data, 235
Qualitative theory, 39
Qualitative variables
  bar graphs and, 401
  measures of association and, 421
  multivariate analysis and, 482–483
Qualitative *versus* quantitative theories, 38–39
Quantitative approach, 235
Quantitive data, 235
Quantitive theory, 38–39
Quasi-experimental designs, 336–339

Quasi-independent variables, 336–339
Questionnaires, 261–268, 274–277

Random assignment, 110–111, 293
Random digit dialing, 281
Random sampling, 165, 278–281
Randomized-multigroup design, 298–299
Randomized two-group design, 294–298
Range, 415
Range effects, 142
R-Square, 479–480
Rating scales, 146
Ratio scale, 137
Rational method, 21–22
Rationalism, 21–22
Raw score regression weight, 424
Reactivity of measurement, 149–156
Realism of a simulation, 122
Reasoning, deductive, 28
Recording
- complexity and, 226–227
- ethnographic data, 240–241
- single events *versus* behavior sequences, 226
- video recording, 227

Recursive models, 494
Refereed journals, 71–73
Reference section, APA style, 518–523
Reflecting strategy, 470
Regression line, 469
Regression weights, 422, 424
Reliability, 130
Repeated-measures designs. *See* Within-subjects designs
Replication,
- direct, 369
- intersubject, 361, 369
- intrasubject, 360, 361
- systematic, 369
- *versus* inferential statistics, 461–463

Representative sample, 277–278
Research
- computer simulations, 194
- defined, 7
- design, 28
- fraud in, 214–220
- *in vitro* methods, 194
- learning about, 6, 7
- steps in, 26–29
- theory-driven *versus* data driven, 51–53
- what is and, 2–6

Research designs
- choosing, 102
- correlational, 104–108
- experimental, 108–114
- experiments *versus* demonstrations, 113–114
- external validity, 118–119
- field setting, 123–124, 169–170
- functions of, 102
- internal validity, 114–120
- laboratory setting, 120–121

Research ideas, sources of, 57–75
Research issues, animal, 190–193
Research process, 26–29
Research question development, 63–66
Research, reporting
- ethics of, 542
- poster sessions, 74–75, 542–543
- presentations, oral, 540–541
- presentations, paper, 74, 540–543
- publishing, 529–540

Research reports, reading, 83–88
Research reports, writing in APA style
- abstract, 505–506
- APA style, 500–501
- author name, 504–505
- author note, 505
- avoiding biased language, 530–531
- citing quoted material, 527, 529
- citing references, 527–527, 528
- commonly misused words, 537
- discussion section, 517
- elements of APA style, 526–538
- expression and organization, 531–539
- figures, 524–525
- footnotes, 522–523
- formatting, 502
- heading structure, 502–503
- introduction, 506–510
- method section, 510–513
- page header, 502–503
- parts and order of, 503–525
- reference section, 518–523
- results section, 513–516
- tables, 523–524
- title page, 504–505

Research settings, 120–124
Resistant measures, 409
Response bias, 384
Responsible Conduct of Research (RCR) program, 218
Restricted items, 262–263
Results section, APA style, 513–516
Reversal strategy, 361
Rho, Spearman's (ρ), 421
Rival hypotheses, 114
Role attitude cues, 150–151
Role playing, 181–182
Rotation of factors, 474–475
Roy-Bargman stepdown analysis, 487
Running head, APA style, 502

Sample
- biased, 269
- defined, 163
- economic, 276
- nonrandom, 165–168
- random, 165, 278–281
- representative, 269
- size of, 276–278, 431, 433, 457
- *versus* population, 165

Sampling
- cluster, 284–285
- distribution, 430–431
- error, 286, 431
- nonrandom, 165–166
- proportionate, 282
- random, 165, 278–281
- simple random, 279
- stratified, 281
- systematic, 283
- techniques, 278–283

Sampling distribution of the mean, 430–431
Scales of measurement, 135–138
Scatter plots, 404
Science
- as a way of thinking, 3
- defined, 2
- pseudoscience *versus*, 9–10
- theory's role in, 43–46

Scientific explanations,
- defined, 11
- failures of, 17–20
- characteristics of 12–13
- *versus* commonsense explanations, 17–19
- *versus* pseudoexplanations, 19–20

Scientific method, 22–26
Scientific theory, 32–34
Scientist, 3
Scope, 43
Second-order meta-analysis, 250
Secondary source, 69–70
Self-report measures, 146
Semipartial correlation (part correlation), 476, 477–478
Serials Index, library, 83
Significance, level of, 459–460
Significance of contribution, 92–93
Simple main effects, 321
Simple random sampling, 279–281
Simple regression analysis, 479
Simulations, 121–122
Single-blind technique, 155
Single-factor designs,
  baseline, 372–376
  randomized-groups, 294–298
  within-subjects, 315
Single-subject designs
  baseline, 360–380
  defined, 291
  discrete trials, 383–386
  dynamic, 380–382
  inferential statistics and, 386–388
Skewed distribution, 409
Small-n approach, *See* Single-subject designs
Social desirability effects, 27
Society for Psychological Science, journals of, 73
Sociogram, 243
Sociometry, 243–244
Solomon four-group design, 346–347
Sources of variance, 442
Spearman rank-order correlation (Spearman's rho), 421
Split-half reliability, 132
Split-plot designs. *See* Mixed designs
Squared semipartial correlation, 480
Stability criterion in baseline designs, 360–365
Stacked format, 393
Standard deviation, 415
Standard error of estimate, 424
Standard error of the mean, 431
Standardized regression weights, 424, 480
Statistical analysis, 294
Statistical errors, 431, 434–435
Statistical regression and internal validity, 117
Statistical significance, 89–91, 435–436
Statistical symbols, 515, 516
Statistical tests. *See* Inferential statistics
Statistics. *See* Descriptive statistics; Inferential statistics
Stemplot, 408
Stepwise regression, 479
Stratified sampling, 281
Strong inference, 50
Structural Equation Modeling (SEM), 495–496
Subject mortality, 351–352
Subject pool, 165
Subjects, *See* Animal subjects; Human participants
Subjects subsection, APA style, 510
Subpopulations, 163
Summary sheets, 393, 394, 395, 396
Suppressor variables, 480
Survey research,
  acquiring a sample for, 277–287
  administering questionnaire, 269–273
  assembling questionnaire, 267–268
  assessing questionnaire reliability, 274–276
  assessing questionnaire validity, 276–277
  attitude measurement and, 258–259
  ethical concerns, 260
  observational research *versus*, 259–260
  questionnaire design, 261–268
  writing questionnaire items, 261–266
Survey types,
  face-to-face interviews, 272–273
  group-administered, 271–272
  internet, 270–271
  mail, 269–270
  telephone, 271
Systematic observations
  automating experiments, 157–158
  choosing a scale of measurement, 138–141
  choosing measures, 130–138
  choosing variables, 128–129
  choosing when to observe, 148
  manipulation check, 158–159
  pilot study, 158
  reactive nature of psychological measurement, 149–156
  tailoring measures to participants, 143–144
Systematic replication, 369

t test, 439–440
Tables in manuscript, APA style, 523–524
Tautology, 19
Technical report, 75
Telephone surveys, 271
Test-retest reliability, 131
Testability of a theory, 46–47
Testable explanations, 97
Themed sentences, 534
Theory, scientific
  characteristics of a good, 46–48
  classifying, 38–43
  defined, 32
  roles of, 43–46
  strategies for testing, 48–50
  theory-driven *versus* data-driven research, 51–53
  versus hypothesis, 33–34
  versus law, 34
  versus model, 34–35
Theory, types of
  analogical, 40–41
  descriptive, 39–40
  fundamental, 42–43
  qualitative, 39
  quantitative, 38–39
Thesaurus of Psychological Index Terms, 79
Third-variable problem, 106
Time sampling, 227
Time series designs, 339–340
Title page, APA style, 504–505
Transitional behavior, 364
Transitional sentences, 534
Treatments, 109
Two-tailed test, 437–438
Two-way interaction, 323
Type I error, 435
Type II error, 435

U.S. Department of Agriculture, 213
U. S. Department of Health and Human Services, 205–206

U. S. Office of Research Integrity (ORI), 215
U. S. Public Health Service, 212
Unit length and writing, APA style, 533
Univariate strategy, 466
Unobtrusive observations, 236
Unplanned comparisons (*post hoc* comparisons), 445
Unrecoverable baselines, 370–371
Unstacked format, 392
Unstructured interview, 272
Unweighted means analysis, 447

Validity
- concurrent, 134, 227
- construct, 134, 276
- content, 133, 276
- criterion-related, 134, 277
- ecological, 139–140
- external, 118–119, 165, 176–177, 361
- face, 133
- internal, 114–118, 174–175, 360
- of a measure, 133–134
- of a questionnaire, 276–277
- predictive, 134, 227
- volunteerism and, 171–170, 174–175

Values, influences on research, 96–98
Variables. *See also* Dependent variables; Independent variable
- choosing for a study, 128–130
- confounding, 115–116, 325–327
- correlational research and, 103–108
- criterion, 105, 261, 467
- defined, 23
- demographic, 261
- endogenous, 493
- exogenous, 493
- experimental research, 108–114
- extraneous, 110–111
- identifying, 251–252
- identifying for meta-analysis, 251–252
- internal validity and, 114–116
- latent, 495
- operationally defining, 64–65
- predictor, 105, 261, 467
- quasi-independent, 336–339
- range effects, 142
- scales of measurement and, 137–138

Variance, 415
Variation, partitioning, 442–443
Volunteer bias, 171
Volunteerism
- and external validity, 176–177
- and internal validity, 174–175
- factors affecting decision to volunteer, 171–174
- remedies for volunteerism, 177

Web of Knowledge, 73
Weighted means analysis, 447
Wilcoxon signed ranks test, 456
Within-group variability, 442–443
Within-subjects designs,
- advantages and disadvantages of, 304–306
- carryover effects in, 306–307
- defined, 291
- MANOVA with, 487–488
- matched-groups designs *versus*, 314–315
- types of, 315–317
- when to use, 313–314

Writing APA style, 500–501, 501–503
Wright's rules and path analysis, 494

z test for two proportions, 441

This page intentionally left blank

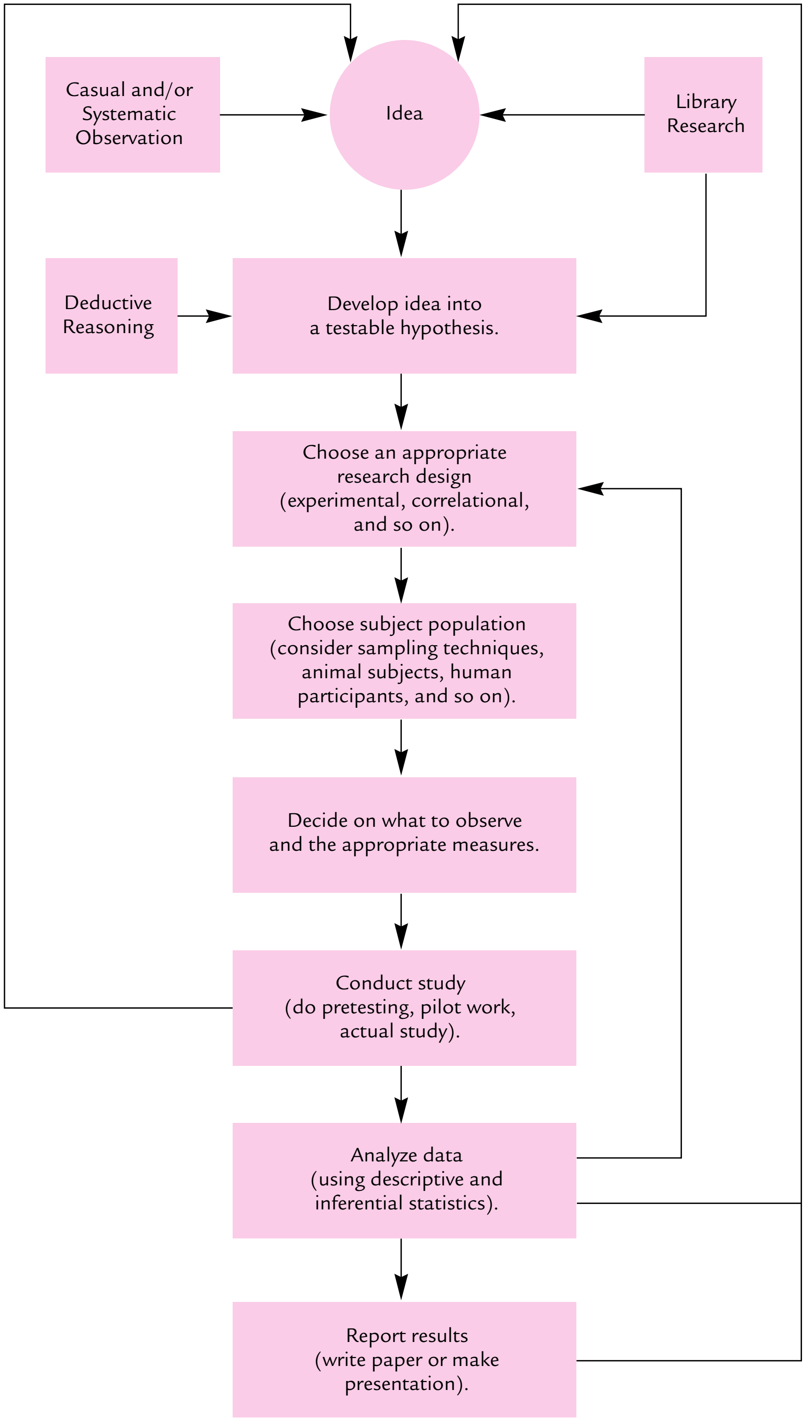

Casual and/or Systematic Observation
Idea
Library Research
Deductive Reasoning
Develop idea into a testable hypothesis.
Choose an appropriate research design (experimental, correlational, and so on).
Choose subject population (consider sampling techniques, animal subjects, human participants, and so on).
Decide on what to observe and the appropriate measures.
Conduct study (do pretesting, pilot work, actual study).
Analyze data (using descriptive and inferential statistics).
Report results (write paper or make presentation).

TABLE 16-5 **Commonly Misused Words**

WORDS	TRUE MEANINGS AND COMMENTS
affect/effect	*affect:* to influence *effect:* the result of; to implement
accept/except	*accept:* to take willingly *except:* excluding; to exclude
among/between	*among:* used when you refer to more than two *between:* used when you refer to only two
amount/number	*amount:* refers to quantity *number:* refers to countable elements
analysis/analyses	*analysis:* singular form *analyses:* plural form
cite/site	*cite:* make reference to *site:* location
datum/data	*datum:* singular form *data:* plural form
every one/everyone	*every one:* each one *everyone:* everybody
few/little	*few:* refers to number *little:* refers to amount
its/it's	*its:* possessive pronoun *it's:* contraction of "it is"
many/much	*many:* refers to countable elements *much:* refers to quantity
principle/principal	*principle:* strongly held belief *principal:* foremost
than/then	*than:* conjunction used when making a comparison *then:* refers to the past in time
that/which	*that:* used to specify a crucial aspect of something: "the study that was conducted by Smith (1984)" *which:* used to offer a qualification that is not crucial to something: "the study, which was published in 1984" (*which* is always preceded by a comma; *that* takes no comma)
there/their/they're	*there:* refers to a place *their:* possessive pronoun *they're:* contraction of "they are"
whose/who's	*whose:* the possessive of "who" *who's:* contraction of "who is"
your/you're	*your:* possessive pronoun *you're:* contraction of "you are"

SOURCE: Compiled from Crews, 1980; Hall, 1979; Leggett, Mead, & Charvat, 1978; and Strunk and White, 1979.