CHAPTER 12 PLANNING THE STUDY

To interrogate nature—that's where the fun is. Carl Sagan

Congratulations! You've made it through the most anxiety-provoking and frustrating phase of the research. Most likely, you have a research question now, and possibly a hypothesis that "makes your heart leap." This and the knowledge you have gained in previous chapters will guide your selection of a research design. Although you still have many other decisions to make about how to proceed in your research, as we will see, the possibilities from which you will be choosing are more limited than they were earlier in the project.

Our goal in this chapter is to help you with the decision making that follows the selection of a design. We will try to anticipate questions you might have on how to recruit and assign participants to conditions; how to write a questionnaire; how to avoid special problems of control that can influence the results of research with human participants; how to write a consent form and apply to your campus institutional review board; and, finally, how to debrief participants at the end of your study.

RECRUITING PARTICIPANTS

Once you have decided on a research design, your first step will be to choose between the two basic methods for recruiting participants, probability sampling and convenience sampling, that were discussed in Chapter 10.

In probability sampling, subjects are selected at random from a population. Typically, this sampling is restricted to surveys, where the researcher wants to generalize the results of the sample to the entire population that was sampled. Probability sampling is time-consuming because all members of the population must be identified and listed, then potential participants must be randomly selected from this list and invited to be in the survey.

Probability sampling is a lot of work compared to convenience sampling, in which any available person is a potential participant. Convenience sampling, typically, is used for selecting subjects for experiments. Of course, the price of this ease of recruiting is that the results of the experiment cannot be generalized with a calculable error to any population.

Once you have decided on your sampling plan, you can implement it using one of the strategies to which we now turn.

PROBABILITY SAMPLING

Chapter 6 described how to draw samples from a population by hand, a time-consuming and difficult process. However, probability sampling can be done more easily with a computer. A web application for random sampling is available at www.muststudy.com/LearnStat/Lessions.

CONVENIENCE SAMPLING

Convenience sampling is more common than probability sampling in student research. In fact, your college may have a ready-made convenience sample, a campus *subject pool,* available for you to use. At many schools, undergraduate majors in psychology are asked to participate in research as part of their course requirements. To avoid the ethical problem of coercing participation, students usually are allowed to choose which studies they will participate in, and offered alternative educational options (like serving as an

observer in research or attending a research presentation; McCord, 1991) if they do not wish to participate. Ask your faculty adviser if there is a subject pool at your school and how it works.

There are alternative methods you can use if your college has no subject pool. One way to recruit participants is to go to a class, where you will reach many potential participants in one face-to-face meeting. Approach a sympathetic faculty member who teaches a large class and ask whether you can take a few minutes at the beginning of class to recruit volunteers for your research. Other methods of contacting volunteers include advertising in the school newspaper, posting notices, asking friends and acquaintances to volunteer, and asking for volunteers in the dining hall or dorms. Campus mailings or telephone contacts are another possibility, but face-to-face recruiting usually works best.

ASSIGNING SUBJECTS TO GROUPS

You will be conducting either an experiment or a correlational study. In a correlational study, the researcher does not set up the conditions of the study, but rather observes the behavior of subjects in naturally occurring conditions. The conditions either are chosen by the subjects themselves during the course of their lives (as in the *Consumer Reports* evaluation of psychotherapy [1995], in which people reported on their experiences with therapists of their choice), or they are characteristics of the subjects themselves (as in Sulloway's [1996] study of birth order and creativity). In an experiment, the researcher sets up the conditions and decides which subjects will be observed in them. In the Elkin et al. (1989) evaluation of psychotherapy, for example, the researchers decided which type of therapy each participant received.

Ideally, the assignment of subjects to groups in an experiment should be done randomly. In Chapter 6, we described how to randomly assign subjects to conditions by hand, but the easiest method of randomizing subjects is by computer. A web application for the random assignment of subjects to groups is available at www.muststudy.com/LearnStat/Lessions.

DECIDING ON APPARATUS AND MEASURING INSTRUMENTS

Special equipment may be needed to precisely control the stimuli you present to subjects, to time their responses, or to record their behavior. The next step in planning your research is to decide on the apparatus and measuring instruments that you will use.

Today we have measuring instruments for research never dreamed of by psychology's pioneers—machines for magnetic resonance imaging, the lie detector, and audio and video recorders that allow researchers to record and play back at different speeds, to name a few. Personal computers now replace many of the instruments used by researchers in the past, and commercial software is available for conducting many types of experiments in perception, memory, and cognition (see Brooks/Cole Research Methods and Statistics Catalog, 1997).

Although B. F. Skinner (1959) had to personally design and build the first Skinner box to control the delivery of food pellets to his animal subjects (see Chapter 9), researchers today can purchase one. In fact, commercial firms now manufacture many of psychology's most commonly used instruments. Your psychology department may even own some of them. Unfortunately, scientific instruments are very expensive, a fact that may limit the kinds of measures

available to you for your study.

You can get valuable information on the instruments to use in your research by reading the literature on the problem of interest to you. If any of the studies you discover used an apparatus specifically designed for the research, the published report usually will include a detailed diagram with measurements; if not, you may be able to get plans for the apparatus from the researcher.

Often the stimulus materials for research are verbal materials, (e.g., a description of an event, a story, pictures) that you can create yourself or get by writing to the originator. Don't be shy! Sharing is an agreed upon ethic among scientists, who want their work to be replicated and extended by others.

Self-report Measures

Because self-report paper-and-pencil measures are used for so many purposes in psychological research (e.g., assessing mood states, attitudes, abilities, and interests), most likely your study will include such a measure. If so, there are three general sources for you to explore: (1) publishing houses that sell commercial tests, (2) scientific journals that publish measures developed by researchers for their own studies, and (3) your own creativity.

Finding commercial tests. Some of the best measures are commercial. Usually extensive information is available on the reliability and validity of such measures. The Buros Institute of Mental Measurement at the University of Nebraska, Lincoln, provides excellent reference material on commercial tests. Box 1 discusses their publications, including the extensive information you will find at their Website on the Internet.

Commercial tests range in price from about twenty dollars for simple paper-and-pencil tests to several hundred dollars for intelligence tests and batteries of achievement and aptitude tests. If you want to use a commercial test in your study, most likely your professor will have to order it for you, or cosign the order form, because there are restrictions on who can purchase such measures. For example, The Psychological Corporation, one of the major test publishers, classifies tests into three categories according to the credentials required for purchasing them. The Wechsler intelligence scales, for instance, are Class C tests, which require a Ph.D. in psychology or education, or verification of required training or experience, to purchase them (The Psychological Corporation, 1997).

BOX 1 FINDING AND EVALUATING COMMERCIAL TESTS: THE BUROS INSTITUTE OF MENTAL MEASUREMENT

The Buros Institute, founded in 1939 by Oscar Buros, provides evaluative information on commercial tests and measurement issues for professionals. *The Mental Measurements Yearbook (MMY),* published beginning in 1938, contains descriptive material on tests, reviews, and references. The volume now is issued on alternate years with *The Supplement to the Mental Measurements Yearbook*. Its companion volume, Tests *in Print,* is a bibliography of all commercial tests in print, and serves as an index for *MMY* reviews. These reference books should be available at your college library.

The Buros Institute's Website at the University of Nebraska, Lincoln, (www.unl.edu/buros) is a treasure-trove of information on tests. The site provides instructions on using *MMY* and Tests *in Print,* and offers several databases that can be searched from the Website for reviews of commercial tests, publishers, and unpublished tests.

The *Test Review Locator* lists the volumes of *MMY* and *Test Critiques,* a reference work on tests published by PRO-ED publishing company, that include reviews of particular tests. To use the *Locator,* you enter identifying information about the test. For example, if you enter "Beck" in the *Locator,* it shows that the *Beck Depression Inventory* was reviewed in the 11th *MMY* and in Volume II, 1985, of *Test Critiques*. Your library should have these volumes, so you can read the reviews. (If not, the institute has a fee based fax service for reviews.) The *Locator* also can be used to find tests you are not familiar with. If you enter "creativity," for example, the *Locator* lists reviews of creativity measures.

Students who wish to purchase a test from this publisher must send them a written request to use the test along with a letter from a faculty sponsor endorsing the student project. Because publishers' requirements vary, you should consult their catalogues to learn what they require. The addresses and telephone numbers of over 900 test publishers can be accessed using the *BUROS/ERIC Test Publisher Locator* at the Buros Institute Website (see Box 3 for the institute's address).

Finding unpublished tests. If you decide not to purchase a commercial test, you may find a suitable "unpublished test" in the psychological literature. Tests published in research articles are called "unpublished" to distinguish them from commercial tests. Box 4 provides information on a variety of directories and books on unpublished tests.

The primary sources for the directories and collections of tests presented in Box 2 were scientific journals, a source that you can explore yourself. In fact, your own search may turn up better possibilities than using the references in Box 2, because you can examine the latest journals using the unique search criteria most appropriate for your study. If you are able to find an unpublished test that you would like to use, the APA recommends that you contact its author to request permission. There is an excellent guide to finding commercial and noncommercial tests published by the APA at their Website (go to www.apa.org, then click on "Science Information"), which ends with this recommendation. Written permission is required for copyrighted tests.

BOX 2 FINDING AND EVALUATING UNPUBLISHED TESTS: DIRECTORIES AND DATABASES PROVIDING INFORMATION ABOUT MEASURES AND THEIR SOURCES

The APA's six volume *Directory of Unpublished Experimental Mental Measures* covers about 6,000 tests, published between 1970 and 1990 (Goldman, Saunders, & Busch, 1996; Goldman, Osborne, & Mitchell, 1996; Goldman & Mitchell, 1995). The entry for each test (covering topics from altruism to zygosity) includes its purpose, format, and reliability, as well as references to studies that have used it. The sixth volume includes a subject index for all the volumes. This series, started in 1974, supplements the *Mental Measurements Yearbooks,* by covering tests not commercially available. The volumes are intended to promote the use of promising unpublished tests, and to make tests available to student researchers who may not be able to afford commercial tests.

Carol Beere's two companion volumes, *Women and Women's Issues: A Handbook of Tests and Measures* (1979) and *Gender Roles: A Handbook of Tests and Measures* (1990), cover inventories related to women's issues and gender roles published between 1927 and 1988. Beere compiled these volumes to encourage gender-related research by making it easier for interested researchers to find quality measures. The latter volume, which covers 211

measures, provides a full description of each measure, as well as information on its reliability and validity, references to research that used it, and a bibliography of related articles.

The Buros Institute's Website (see Box 3) includes the Educational Testing Service (ETS)/ *ERIC* test collection, which can be searched for information on over 10,000 tests. *Health and Psychosocial Instruments,* put out by Behavior Measurement Database Services (www.ovid.com), another extensive database on tests, covers over 15,000 tests. This database is available through libraries and on CD-ROM.

Reference works containing tests and inventories:

Robinson et al.'s three volumes, *Measures of Political Attitudes* (Robinson, Rusk, & Head, 1973), *Measures of Occupational Attitudes and Occupational Characteristics* (Robinson, Athanasiou, & Head, 1973), and *Measures of Social Psychological Attitudes* (Robinson & Shaver, 1973), include a variety of inventories.

Measures for Clinical Practice: A Sourcebook (2nd ed.): Vol. 1. *Couples, Families and Children,* and *Measures for Clinical Practice: A Sourcebook* (2nd ed.): Vol. 2. *Adults* (Fisher & Corcoran, 1994) are two handbooks containing scales for use in clinical practice with children, adults, and families.

Family Assessment: Inventories for Research and Practice (McCubbin & Thompson, 1991) covers family assessment inventories.

Behavior Analysis Forms for Clinical Intervention and *Behavior Analysis Forms for Clinical Intervention* (Vols. 1, 2) (Cautela, 1977,1981) present forms for use in behavioral-clinical interventions.

ETS has a collection of tests on microfiche that is available at libraries or directly from ETS. The ETS test collection is described at the Buros Website (see Box 3).

The Keirsey Temperament Sorter, a measure of Jungian personality types, can be taken, scored, and interpreted at http://sunsite.unc.edu/jembin/mb.pl.

CONSTRUCTING YOUR OWN QUESTIONNAIRE

You may decide to write your own questionnaire rather than using one developed by someone else. With your own measure you can get at the precise distinctions you want to make. The tips from the experts that we offer in this section are designed to help you make decisions on the format and wording of your questions.

Open versus Closed Questions

Your first decision in developing your own questionnaire will be which of the two basic types of questions to use—*open-ended* or *closed*. Open-ended questions do not impose restrictions on participants' responses; closed questions do.

Open-ended questions require participants to construct their own answers. *Closed questions* require them to choose answers from a list of options.

The following question is open-ended:

What products do you think the Federal Drug Administration (FDA) should regulate?

Asked as a closed question, it might read:

Please check the products you think the Federal Drug Administration (FDA) should regulate.

- □ prescription drugs
- □ herbal medicine
- □ cigarettes
- □ vitamins
- □ toothpaste

A combination question could be created by adding an open-ended alternative (e.g., "other ") to this list. Questions of any of these types can be in cluded in the same questionnaire.

Open-ended questions are necessary when the response alternatives are too numerous to list, for example, when participants are asked to report their occupations or majors. They also are useful in the early stages of research when investigators may be unsure of the range of possible answers to their questions. Because participants can answer open-ended questions in unanticipated ways, this format also affords researchers an opportunity to learn new things. In case studies, participant observation, and phenomenological studies, open-ended questions capture the richness of people's experience in ways that closed questions, with restricted response alternatives, cannot.

Once the variety of responses is understood, forced-choice questions offer researchers the advantages of standardizing the options from which respondents may choose and ease of item scoring. With open-ended questions, participants' answers may be ambiguous and they may not think of answers that are critical for the study, like the possibility of the FDA regulating cigarettes.

Closed Item Formats

Box 5 illustrates several common formats for closed questions as well as two unusual ones. For each of these, we give several examples of how the format would appear in test items. The items from commercial tests in the box are written in the same style as the items in the tests but are not the actual test items.

Formats 1 and 2, both variants of the agree-disagree, true-false format, are common in personality inventories. Format 2 is a variant of Format 1, which includes "don't know" or "not applicable" among the response alternatives.

Rating scales, the third format in Box 5, are useful for measuring the intensity of behaviors, opinions, or moods. This type of measure is called a Likert-type scale, after Rensis Likert, the scientist who introduced it for attitude measurement (Likert, 1932). Box 5 includes four-, seven-, and one-hundred-point scales for rating intensity. A typical questionnaire written in this format would include several questions on the same topic with the same rating alternatives. The scores on the separate items would be added to yield a total score.

BOX 5 QUESTION FORMATS FOR CLOSED ITEMS

Format 1. Agree-Disagree Items

Minnesota Multiphasic Personality Inventory-2[1] (MMPI-2), used for clinical diagnosis of patients.

T F Sometimes I see myself brushing my teeth with the wrong hand.

Survey question on politics (Schuman & Presser, 1981).

Do you agree or disagree that: Most men are better suited emotionally for politics than are most women.

[ ] Agree [ ] Disagree

Coopersmith Self-esteem Inventory[1].

Like Me □ Unlike Me □ I do not worry too much about things.

Format 2. Agree-Disagree, with a "Don't Know" Alternative

Cattell 16 PF[1] (Primary Factors), a personality test which measures 16 basic personality traits.

I worry too much about the future.

a. Hardly ever

b. ?

c. Often

Strong Vocational Interest Blank[1], an interest inventory used in vocational counseling.

L: Like I: Indifferent D: Dislike

LID Riding a mountain bike down a steep **hill**

Format 3. Rating Scales

Hare Self-esteem Scale, a self-report measure for school age children (Hare, 1985). a = Strongly disagree b = Disagree c = Agree d = Strongly agree

1. I have at least as many friends as other people my age.

Beck Depression Inventory[1], a self-report measure of the severity of depression (Beck & Steer, 1987).

Tomorrow will be a good day.

I am not looking forward to tomorrow.

Nothing good will happen tomorrow.

Tomorrow will be as disappointing as every other day.

Bern Inventory[1], a measure of gender roles (Bern, 1978). Indicate how true of you the following characteristic is: good listener

Never or almost never true	Usually not true	Sometimes but infrequently true	Occasionally true	Often true	Usually true	Always or almost always true

Family Sense of Coherence Scale (Antonovsky & Sourani, 1988). To what extent does it seem to you that family rules are clear?

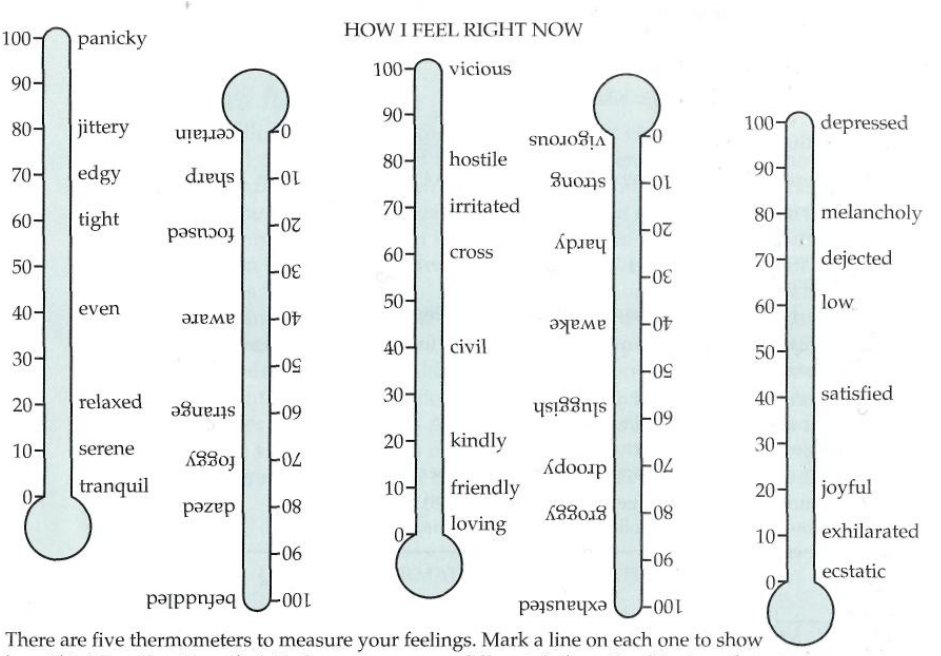
1	2	3	4	5	6	7
The rules in the family						The rules aren't clear at all.
are completely clear						

Semantic Differential, an instrument used to measure the meaning of abstract concepts (Osgood, Suci, & Tannenbaum, 1957).

My Spouse

Strong : : : : : : Weak

Mood Scales (Tuckman, 1988)



There are five thermometers to measure your feelings. Mark a line on each one to show how "high" or "low" you feel. Each one measures a different feeling. Don't just mark them all the same. For two of them, you have to turn the paper over. Give your real, honest feeling. Don't just make something up.

Format 4. Forced Choice Question

Political survey question (Schuman & Presser, 1981). Would you say that

- ☐ most men are better suited emotionally for politics than are most women,
- ☐ that men and women are equally suited,
- ☐ that women are better suited than men in this area?

Format 5. Rank Order Questions

The subject ranks the response alternatives according to their preference. How important are each of the following in buying a new car?

Performance

Reliability

Style

Price

Size

Manufactured in US

Quality

Format 6. Magnitude Estimation (Lodge, 1981; in Converse & Presser, 1986)

I would like to ask your opinion about how serious YOU think certain crimes are. The first situation is, "A person steals a bicycle parked on the street." This has been given a score of 10 to show its seriousness. Use this situation to judge all others. For example, if you think a situation is 20 TIMES MORE serious than the bicycle theft, the number you tell me should be around 200, or if you think it is HALF AS SERIOUS, the number you tell me should be around 5, and so on.

COMPARED TO THE BICYCLE THEFT AT SCORE 10, HOW SERIOUS IS:

A parent beats his young child with his fists. The child requires hospitalization.

A person plants a bomb in a public building. The bomb explodes and 20 people are killed.

Format 7. Randomized Response

I would like to ask you whether you have ever used marijuana, but I don't want you to answer directly because it is illegal to use marijuana. So I will ask you to follow a procedure that will make it safe for you to answer. After you answer, no one will know if you have used marijuana or not, but from all the answers in the school I will be able to estimate the percent of students that have tried marijuana. I want you to answer the question differently depending on the number you get from the number target. Hold your finger above the circle below, then shut your eyes and place your finger on the circle. Don't tell anyone the number you picked.

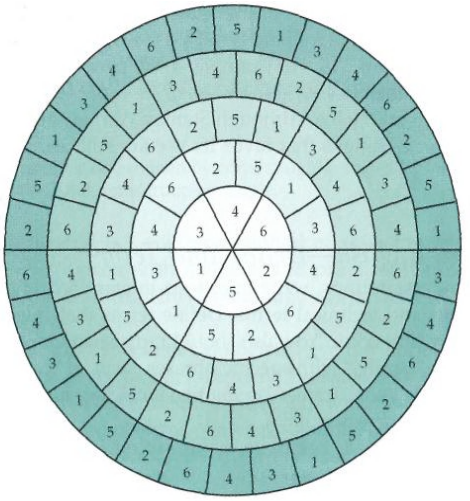
Now, if you got a **1, answer No** to the question.

If you got a **2, answer Yes** to the question.

If you got 3, **4, 5, or 6, then answer truthfully Yes or No** to the question:

Have you tried marijuana?

Answer: Yes No



'This item and the other items in this box from commercial tests are written in the same style as the items on the tests but are not actual items from the tests. [2]The number target is from Reaser, Hartsock, & Hoehn (1975).

Format 4, the forced-choice question, requires respondents to select their answers from several mutually exclusive alternatives covering all possible responses to the question. This format has the advantage of making all the response choices explicit, ensuring that respondents consider each alternative.

Rank-order questions, Format 5, require respondents to order a set of alternatives according to their preferences, a procedure that generates a great deal of information from a single question. Unfortunately, such rankings can be difficult to make, and consequently, unreliable.

The last two formats in Box 5, magnitude estimation and randomized response, are rare in psychological questionnaires, but both are worth considering for some applications.

Format 6, magnitude estimation, might seem impossible at first glance. However, in the example in Box 5, Lodge (1981) asked subjects to assign a number representing the seriousness of killing 20 people with a bomb, as well as other crimes, and they could do so! Magnitude estimation was popularized by S. S. Stevens, the psychophysicist who proposed the classification of scales discussed in Chapter 4.

The randomized response question, Format 7, is used to ensure the confidentiality of participants' answers to sensitive questions, like whether they have tried controlled drugs (Brown & Harding, 1973) or had premarital sex (Krotki & Fox, 1974). The question in Box 5, for example, asks respondents whether they have ever tried marijuana, an illegal drug. The way the question is asked prevents anyone from figuring out from the answer whether a particular subject has or has not tried the drug! But from the answers for the entire group of respondents, the researcher can estimate the *percent* of people in that group who have tried the drug. For this question, the estimate of the proportion of respondents trying marijuana is (6 Py-l)/4, where Py is the proportion of respondents answering yes. For example, if 100 people answer the question and 30 respond yes, then (6 (.30) -l)/4 = .20, or 20%, of the 100

people are estimated to have tried marijuana. Although Stanley Warner introduced this technique in 1965, the conditions that favor its use are still being studied (Fox & Tracy, 1986).

Use Standard English—Define Your Terms

To avoid misunderstandings, write questions in standard English, avoiding slang expressions and technical terms. Define terms that might be misinterpreted when they first occur. You also should replace any terms with a vague reference, such as "family," with specific phrases, such as "people living in the same household" or your "immediate family, including just your spouse and children" (Converse & Presser, 1986).

Once your first draft is complete, try it out on friends. Have them read each question out loud, then paraphrase the question and give their answers to it, explaining what they mean. With this procedure, you should be able to catch ambiguities in wording and discover ways to rewrite questions for greater clarity.

EVALUATING PSYCHOLOGICAL MEASURES

Regardless of the source of your measure (commercial, unpublished, or tailor-made), you will want it to be reliable and valid for your purpose. In this section, we discuss these properties of measures.

Reliability

The *reliability* of a measure is the degree to which repeated measurements of the same subjects under the same conditions yield consistent results.

Reliability is assessed by computing the correlation between the outcomes of two different administrations of a measure to the same group of subjects. A correlation coefficient of zero, $r = 0$, indicates a completely inconsistent measure; a correlation of one, $r = 1$, indicates perfect consistency; values between 0 and 1 indicate the degree of reliability.

First, consider a completely unreliable measure. Imagine that you "measure" a group of people by rolling two dice for each person, assigning them the sum of the two numbers as their scores. The possible values for this measure would range from 2 to 12, If you then "measured" the same people again several weeks later and compared the two sets of scores, you would find no consistency, that is, no reliability. The correlation between the two administrations of this "measure," most likely, would be close to zero.

The Wechsler Adult Intelligence Scale, WAIS-R, is at the other end of the reliability continuum from rolling dice. The WAIS-R manual reports that the correlation between two administrations of the test several weeks apart is $r = .95$ for 25-35-year-olds (Wechsler, 1981). When the same test is administered twice to assess reliability, as in this case, the result is called the *test-retest reliability* of the measure.

Test-retest reliability is determined by correlating the scores for two administrations of the same test to the same group of people.

Although test-retest is a common method of assessing reliability, it does have problems associated with memory and practice. If people remember questions from the first test, this knowledge can affect their scores on the retest.

To avoid such problems, methodologists have developed a procedure for

assessing reliability using parallel forms of the same test. Parallel forms of a test are versions with different item content, but with the same type and difficulty level of items. Parallel forms of a vocabulary test, for example, would include different words of comparable difficulty. The item format (e.g., multiple choice, matching, etc.) would be the same for both forms.

Parallel-forms reliability is assessed by correlating the scores on parallel forms of the test administered to the same group of people at different times.

A disadvantage of using parallel forms is that, just as in test-retest reliability, subjects must be tested twice. In addition, the test developers have to construct alternative forms of the same test, which can be a difficult and time-consuming process. For these reasons, few tests are available in parallel forms.

These problems can be sidestepped, however, by using a clever procedure for determining reliability from the results of a single administration of a test. The subjects first take the test; then the items in the test are divided into halves, so that each half is a short version of the total scale. This creates, in effect, two short parallel forms of the same test. When the measures are psychological tests composed of a series of items, like the WAIS-R, this is easy to do. (With a weight or temperature scale, it is impossible.) Once half the items are assigned to scale "A" and the remaining items to scale *"B,"* the two scales are scored separately, and the results correlated. The resulting correlation is the *split-half reliability*.

The *split-half reliability* of a test is determined by dividing its items into two halves and correlating participants' scores on these parts. The test is administered only once, to one group of people.

For the WAIS-R, the split-half reliability for 20-24-year-olds is .94 (Wechsler, 1981). This reliability is for *half the* WAIS-R test. The reliability for the full test, because it is longer and should give a more consistent result than just half the test, is expected to be higher.

Researchers using the split-half method can estimate full-scale reliability by using the Spearman-Brown formula, which was developed for this purpose. Given the half-scale reliability, this formula computes an estimate of the full scale reliability, assuming that the full scale test is a direct extension of the half-scale test. For the WAIS-R, the formula gives a full-scale reliability of .97, an increase over the split-half reliability of .94. The Spearman-Brown formula is explained by Nunnally and Bernstein (1994) and in other texts on psychological testing.

The advantage of split-half reliability over the test-retest and parallel forms methods is that it requires only one administration of the test. The method also works well for measures that can vary markedly from day to day, like mood. However, the split-half reliability will depend on how the items in the full test are assigned to the two halves; different assignments result in different estimates of the full-scale reliability. This source of error can be overcome by using another method of assessing reliability, alpha reliability (see Nunnally & Bernstein, 1994).

The *alpha reliability* of a scale is equal to the average of all the split-half reliabilities computed for every possible assignment of items to the two halves.

When the scale items have a true-false format, the alpha reliability

sometimes is called the *KR-20 reliability*. There is a special formula for this case.

If you are not using a test with established reliability, you should plan your study so that you can compute the reliability of your test. Computing reliability will establish an important benchmark for your measure: If two versions of your test (the split halves) do not correlate with each other, it is unlikely that the scale will correlate with anything else. If the halves do correlate, the chances are better that the test will correlate with other measures.

Validity

The validity of a measure is concerned with its usefulness.

A test is *valid* to the extent that inferences made from it are appropriate, meaningful, and useful. (Gregory, 1996, p. 107)

Research to evaluate the validity of measures falls into three general categories:

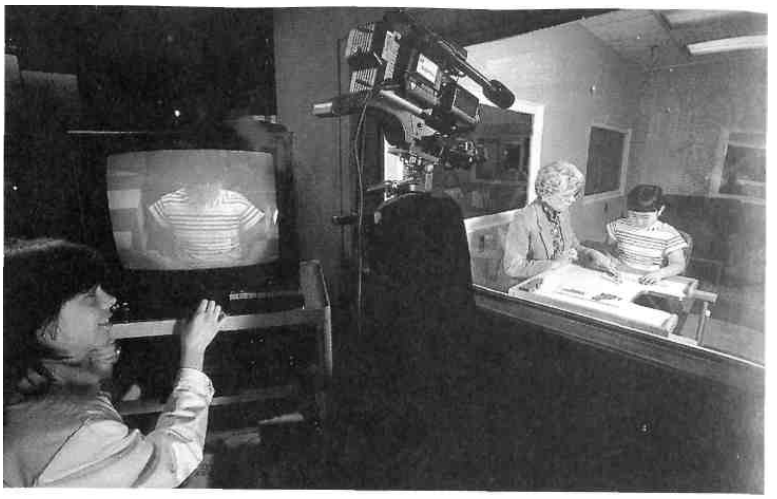
1. Studies on *criterion validity* investigate whether or not test scores predict future behavior or diagnose a present condition. For example, the publishers of the Scholastic Aptitude Test (SAT) claim that the test predicts college grades of students while they are still in high school. Validation here is straightforward. High school students who have taken the SAT are followed through college and their GPAs correlated with their SAT scores.
2. Studies on *content validity* investigate whether questions in a test are a fair and representative sample of the content they are supposed to examine. Experts in the content area of the test usually are involved in these studies.
3. Studies on *construct validity* investigate whether tests are good measures of the psychological concepts their authors claim they measure. For example, a construct validity study might investigate how people scoring high and low on a scale of extroversion behave socially.

A measure can be valid for one application but not for others. As Nunnally and Bernstein (1994), experts on validity, note, "One validates the *use* to which a measuring instrument is put rather than the instrument itself" (p. 84). This is an important idea to keep in mind as you evaluate possible measures for your own research. Pay particular attention to whether any measure you are considering has been shown to be valid for your specific purpose. To do so, look at how the measure has worked in research similar to your own.

EVALUATING OBSERVATIONAL MEASURES

Researchers assess the reliability of observations and tests using similar procedures. To evaluate the reliability of observations, researchers compare the ratings of different observers on the same set of subjects' behaviors.

Inter-observer or *inter-rater reliability* is demonstrated by showing that observers agree in classifying subjects' behaviors.



Videotaping a child's performance through a one-way mirror.

Good inter-observer reliability indicates that the observations are not idiosyncratic, that more than one observer can consistently apply the classification system.

The easiest way to determine inter-observer reliability is to videotape behaviors that are typical of how subjects will behave in the study. The tapes then can be used to train the observers and test for reliability.

Observers should be trained until they demonstrate a high level of agreement with the ratings of an experienced observer. Ninety percent agreement is a common criterion for training. For example, let's say you plan to have observers rate the violence in selected scenes from television shows on a four-point scale. The observers would be trained until their ratings match the ratings of an experienced observer for 90% of the scenes.

To determine reliability, set aside one tape from the videotaped sessions you use to train your observers. Have the observers independently categorize or rate the behaviors on the tape, then assess the extent of their agreement. If you cannot videotape, have two observers rate the same subjects' behaviors simultaneously.

Although correlation is used to establish the reliability of tests, this statistic is not used to assess inter-observer reliability. Recall from Chapter 5 that one reason for the correlation coefficient's popularity is that it allows researchers to compare measures with different units or scales (e.g., foot size and height, creativity and birth order). But in assessing the reliability of observations, we want to know the *exact agreement* of different observers, and the correlation coefficient does not indicate exact agreement.

For this reason, psychologists report inter-observer reliability either as the percent of exact agreement between observers, or by using a related statistic, Cohen's kappa. Percent agreement is calculated by tallying the number of times that two observers give the same rating to the same behaviors, then dividing by the number of ratings, and multiplying by 100, to convert to a percent. Cohen's *kappa* corrects the percent agreement for the possibility that observers could agree at a high level by chance alone. The calculation and interpretation of kappa is discussed in Cohen (1960) and Bakeman and Gottman (1986).

If the observation phase of your study goes on for a long time, you also should check for *intra-observer reliability*.

Intra-observer reliability is the agreement in the ratings of the same observer at different times. Such reliability establishes that observers are consistent in their ratings over time.

Intra-observer reliability is checked by having the observer rate the same recorded behaviors at different times. If the results show insufficient agreement, the observer will have to be retrained.

In most studies, there is no issue about the validity of the observations. Observations often have a direct and immediate interpretation not typical of test scores. If a reliable observer records the occurrence of an event, such as the use of a tool by a chimpanzee, it is assumed that this is a valid observation. Research on the validity of observations usually is not necessary.

SPECIAL PROBLEMS OF CONTROL WITH HUMAN PARTICIPANTS

In previous chapters of this book, we have presented several examples of the special problems faced by researchers studying people. When they investigated Mesmer's animal magnetism, Franklin's commissioners discovered people who believed they had been cured by a treatment that later turned out to depend on suggestion. In Chapter 9, we saw that teachers who believed that facilitated communication would help their autistic clients express themselves began to control what their clients typed, without any awareness that they were doing so. In Chapter 10, we discussed the problem of subject reactivity, the possibility that being observed, per se, may alter subjects' behaviors in a study. In the next sections, we look more closely at such special problems and make recommendations on how you can control them in your own research.

The Hawthorne Effect

Social scientists began to recognize reactivity as a research problem early in this century when a study on worker productivity that began in the 1920s at the Hawthorne plant of the Western Electric Company, a telephone assembly factory, was published. The researchers (Roethlisberger & Dickson, 1939) wanted to measure workers' productivity, selecting the number of telephone relays assembled in a given unit of time as their measure. They decided to separate a small group of the plant's employees from the rest of the workforce, to enable them to record their behavior more accurately, to gain better control over extraneous events, and to prevent general disruption.

Workers in the relay assembly test room at the Hawthorne Works.

They first measured worker productivity in the regular shop, then in the new test room, and again when the company introduced a change in pay. Following this, the researchers introduced a series of improvements in the work situation, including changes in the length and number of rest periods, and in the length of the work week. At one point the workers even received a complimentary lunch during the morning break. Toward the end of the experiment, the original work conditions of no lunch and no rest periods were reinstated.

To their surprise, the authors reported that none of these changes was related in a one-to-one fashion to average hourly productivity, which continued to rise throughout the course of the study, even when the improvements were taken away. In fact, contrary to expectation, when the rest periods and complimentary lunches gradually were eliminated, total weekly output continued to increase, reaching an all-time high level when they were gone. On top of all this, the mental attitude of the workers improved.

The researchers concluded that their findings were a consequence of the special social circumstances created for workers in the experiment. Unlike regular shop workers, the employees in the test room were told of planned changes in the work situation in advance and asked about their opinions, fears, and concerns; they got rests and lunches not offered to regular employees, and they were allowed to talk as they worked; top management was interested in their progress; and their physical and mental well-being were concerns of the investigators. As a result, the researchers concluded, a cohesive social group had come into being, leading to the observed increases in motivation, productivity, and morale.

Roethlisberger and Dickson's interpretation of their results was challenged later by other investigators. Their reanalyses of the data collected at the Hawthorne works have shown that, in fact, productivity did not consistently increase throughout the study and the workers were not a cohesive group, happy with their special work conditions (Parsons, 1974; Bramel & Friend, 1981; and Rice, 1982). Nevertheless,

> psychologists continue to refer to the reactivity of people to the special treatment and attention they receive as research participants as the *Hawthorne effect*.

Demand Characteristics and Experimenter Expectancies

The special problems of studying people became the focus of research once again in 1962 when Martin Orne published one of the first papers to study the experiment as a social situation. Orne's research showed that people behave differently in experiments than they do in other situations. In experiments, they willingly perform dull, meaningless tasks for hours on end and engage in dangerous, even potentially lethal acts that they would never consider doing outside of an experiment (e.g., handling deadly snakes, or putting their fingers into corrosive acid). How, Orne wondered, could such seemingly bizarre behavior be explained? His answer focused on the attitudes we learn about science and scientists in this culture.

As a society, Orne argued, we hold scientists and their work in high esteem. We learn early that scientific research is essential, that it leads to important benefits. Because of these beliefs, participants come to experiments ready to assume the role of "good experimental subject," to put themselves in the

experimenter's hands, much as a hypnotic subject might, ready to willingly perform tasks assigned to them, no matter how boring, uncomfortable, or painful they might be. They concoct purposes for meaningless tasks and trust that experimenters will not let harm befall them. To be "good subjects," they believe, they must cooperate, not "ruin the experiment," and help experimenters find what they are looking for. The experiment becomes a special problem-solving situation in which good subjects develop interpretations of researchers' purposes using any cues that might reveal this.

The cues that suggest hypotheses to participants are called *demand characteristics* (Orne, 1962).

Demand characteristics are present in campus talk about the experiment, in details of the research setting, and even in the experimental design itself. If demand characteristics suggest particular behaviors to participants, and if they are motivated to comply, it follows that the effects of demand characteristics possibly might be mistaken for the effects of independent variables. To test this idea, Orne and Scheibe (1964) conducted an experiment to find out whether some of the effects usually attributed to sensory deprivation actually might result from demand characteristics.

The participants in sensory deprivation experiments usually are isolated in a testing room, where visual, auditory, and kinesthetic stimulation are severely reduced. Because sensory deprivation is disturbing and has been shown to result in disruption of intellectual functioning and in abnormal behavior,

A participant in a standard sensory deprivation experiment.

participants in such research are required to undergo physical examinations and sign release forms before they begin. During the experiment,

participants may be asked to report any strange experiences they have; there may even be a "panic button" in the testing room for them to press should they become too uncomfortable to continue.

Orne and Scheibe believed that some of the usual effects of sensory deprivation might be due to such special features of the experimental setting. To test this hypothesis, they created a "meaning deprivation" group whose members were exposed to such demand characteristics in the absence of sensory deprivation. Even though they would undergo no sensory deprivation, the people in this group were asked to report on their medical history and shown an "emergency tray" filled with drugs and medical paraphernalia that would be available in the test room for their safety. After this, the participants were taken to a "well-lighted cubicle containing 'food and water'," and given "an optional task" to keep them occupied for the four hours of the study. They received no other information about the purpose of the experiment. The people in the control group, who were treated identically to those in the "meaning deprivation" condition, were told that they were control subjects in a study of sensory deprivation. The results of the study supported Orne's hypothesis that participants' responses to demand characteristics can significantly affect research results.

Robert Rosenthal (1994), whose work complements Orne's, has argued that experimenters also unwittingly contribute to the invalidity of research by allowing their expectations to influence their findings. Rosenthal's research suggests that experimenters may inadvertently communicate their expectancies, or hypotheses, to research participants; if Orne is correct, they, in turn, use these cues to tell them how to behave in the study. In research on people, then, *experimenter expectancy effects* can be considered as one type of demand characteristic.

Rosenthal's first studies demonstrated that experimenters' expectations of how their subjects would rate photographs of people were related to the ratings they actually received (Rosenthal & Fode, 1961; as cited in Rosenthal, 1994). Since then, experimenter expectancy effects have been demonstrated in many types of research, including animal learning, person perception, and reaction time studies (Rosenthal, 1994).

Controlling for Suggestion and Reactivity

Demand characteristics, experimenter expectancy effects, and the Hawthorne effect can be controlled by using the same sorts of experimental techniques for control that we have discussed throughout this book, namely: (1) control by holding events constant, (2) randomization, and (3) statistical control.

Holding events constant. First, to the extent possible, researchers should take pains to give the same demand characteristics, as well as the same degree and kind of attention, to the experimental and control groups. Instructions to participants might be prerecorded on audio or videotape to avoid even unconscious bias. "No treatment" control groups, that receive no attention and, consequently, the expectation that their behavior will not change, should be avoided. Instead, control subjects should be given "rival treatments" that hold some promise of being effective and involve the same type of relationship with the experimenter as the experimental subjects have. In addition, whenever possible, experimenters, observers, and participants should be "blinded" as to the subjects' experimental conditions to eliminate differential effects of experimenter expectancies and demand characteristics on

them.

Randomization. We have discussed the importance of randomly assigning subjects to experimental treatments many times in this book. If there are many observers or experimenters in a given study, they also should be randomly assigned to the conditions of the study. Such randomization will avoid any systematic bias that could arise if such people were assigned to observe and administer the treatments in the study in some other way.

Statistical control. If demand characteristics cannot be controlled, consider assessing the impact of these cues on the results by means of a postexperimental inquiry (Orne, 1962). You could interview the participants at the end of the experiment to find out whether they were aware of your hypothesis during the experiment (see the discussion of debriefing later in the chapter). If some were, you then could compare the experimental behaviors of those who were and were not aware of the actual hypothesis and correct for the effects of such knowledge by statistical methods.

Finally, for a given research topic, there may be no need for any special procedures for assessing the effects of demand characteristics or experimenter expectancies, because there may be no logical way that the presence of demand characteristics could explain the results (e.g., when participants report their sensory experiences in relation to slight variations in stimulation; in many learning experiments; in studies of infant behavior at various stages of development, etc.). In fact, Orne concluded:

> The need to concern oneself with these issues becomes more pronounced when investigating the effects of various interventions such as drugs, psychotherapy, hypnosis, sensory deprivation, conditioning of physiological responses, etc., on performance or experiential parameters . . . or . . . where attitude changes rather than performance changes are explored. (Orne, 1962, p. 156)

DEBRIEFING

The term *debriefing,* which originated in the military, has several meanings, all of which apply to how debriefing is used in research with human participants. According to the *Random House Unabridged Dictionary* (Flexner, 1993), the first two meanings of the verb "to debrief" are:

1. To interrogate (a soldier, astronaut) on return from a mission in order to assess the conduct and results of the mission.
2. To question formally and systematically in order to obtain useful intelligence or information.

One purpose of debriefing in research is to provide investigators with useful information on how participants understood the experiment's purpose and behaved during the experiment. In drug studies, for example, debriefing can tell researchers whether the participants complied with the recommended doses of drugs, or whether the people who received placebos guessed that their medication was inactive. This information on participants' expectations and behavior during the experiment can be used in analyzing the results.

The third dictionary definition of debriefing has to do with cautioning people involved in special operations against revealing privileged information to others.

3. To subject to prohibitions against revealing or discussing classified information, as upon separation from a position of military or political

sensitivity.

This same sort of caution can be applied in the research setting. In a debriefing session, participants can be asked not to discuss the experiment with other potential subjects until the study is complete. This is important because prior information about the experimental procedures can affect how people behave in many studies. In fact, Marans (1988) advises experimenters to ask participants to sign a nondisclosure statement during the debriefing session and to take it home with them, as a reminder not to talk about the study until a later date.

The final meaning of the term "to debrief" originated in psychology:

4. *Psychol,* (after an experiment) to disclose to the subject the purpose of the experiment and any reasons for deception or manipulation.

The APA's ethical standards require researchers to debrief participants if they have been deceived in the research. Debriefing allows researchers to correct any misconceptions that they create as part of the study.

There is one final reason for debriefing that is not mentioned in the dictionary definition. Recall from Chapter 7 that the Belmont Report's principle of beneficence requires researchers to do everything possible to maximize the benefits people gain from their participation. One such benefit is educational. During the debriefing session, you can educate participants on the purpose of the research, current knowledge in the field related to the research question, and what you have learned from conducting the study. Such information will give subjects a sense of their role in advancing knowledge, something that most people will feel good about.

APPLYING TO THE INSTITUTIONAL REVIEW\

Most likely, you will have to apply to the institutional review board (IRB) at your college to review the ethics of your proposed research before you begin collecting data. Even if your school does not have an IRB, you should be aware of the concerns of IRBs so that you can plan a study in conformity with accepted ethical practice.

The steps to take and the forms to complete for an IRB review will differ somewhat from school to school, so make sure you know the procedures at your college. If you plan to include people from another institution (e.g., a college, day care center, or clinic) in your research, you also will have to get approval from the IRB of that institution before proceeding. IRBs that review psychological research operate according to the ethical principles of the Belmont Report, federal and state law and regulations governing research with human subjects, and the APA ethical principles. So review the material in Chapter 7 and read the Belmont Report, reprinted in Appendix A, before completing the IRB application.

Recruitment Procedures

According to the Belmont Report's *principle of justice,* people from all walks of life should share equally in the burdens and benefits of research. Accordingly, you should invite as diverse a cross section of people to participate in your study as is permitted by your research design. In your application to the IRB, describe your sampling scheme (see Chapter 10) and how you plan to recruit participants. If you intend to use a probability sample, you should describe the sampling procedure. If your sampling will be done by convenience, describe where and how you will recruit participants. Provide

enough detail so that the members of the IRB will be able to judge whether your research is equitable in its recruitment procedures.

The *principle of respect* requires that potential subjects in research decide for themselves whether to become involved, with full knowledge of what their participation will entail. Accordingly, your invitation to them must be straightforward, describing the procedures, benefits, and risks, if any, as completely as possible. There must be no coercion of any sort, no threats of retaliation or loss for failure to participate, and no remuneration out of proportion to the requirements of the study. To ensure compliance with this principle, the APA guidelines require participants to sign a consent form prior to taking part in research.

The Consent Form

According to Joan Sieber, author of a book to guide students and IRBs through the review process, the consent form should describe the research and its purpose in simple, nonscientific language that is both friendly and respectful of potential participants. An appropriate consent form, in Sieber's view, should include the following points, reprinted from her book (Sieber, 1992, p. 35):

1. Identification of the researcher.
2. Explanation of the purpose of the study.
3. Request for participation, mentioning right to withdraw at any time with impunity.
4. Explanation of research method.
5. Duration of research participation.
6. A description of how confidentiality will be maintained.
7. Mention of the subject's right of refusal without penalty.
8. Mention of the right to withdraw own data at end of session.
9. Explanation of any risks.
10. Description of any feedback and benefits to subjects.
11. Information on how to contact the person designated to answer questions about subjects' rights or injuries.
12. Indication that subjects may keep a copy of the consent.

Remember that special precautions must be taken when potential participants include people who cannot be expected to give their informed consent (e.g., children, prisoners, and mentally disadvantaged people; see Chapter 7). Also, for certain types of research, you may ask the IRB to drop the requirement of informed consent, for example, when the research involves observing public behavior, when answering your questionnaire is tantamount to giving consent, and in extraordinary cases when consent would be impossible, like the emergency room study described in Chapter 10.

Box 4 reprints a sample consent form from Sieber's book that you can use as a model in writing your own. Remember to include a copy of the consent form in your application to the IRB.

BOX 4 SAMPLE CONSENT FORM (REPRINTED FROM SIEBER, 1992)

(Letterhead of the Researcher's Institution) Dear

Patient,

I am a psychologist who specializes in the study of taste perception. I am currently working with the staff of your department to see if we can learn ways to enhance your enjoyment of the food served to you here. We need your help in a new study on how sensitive people are to different tastes and which tastes they prefer. The results of this study may help doctors and dietitians, here and at other hospitals, plan diets to improve health, and may add to the understanding of taste perception.

In this study, we will find out how readily persons detect and identify sweet, sour, salty, and bitter tastes, and which tastes are preferred. This information will be analyzed in relation to some information that I am given by the staff physician from participants' medical records about their age, sex, smoking history, duration of lithium administration, and current lithium concentration. Persons participating in this study can expect to spend about 20 minutes on each of five different days. Participants will be asked to taste plain water and samples of water mixed with small amounts of some safe substances that normally are used to season food; they will be asked to answer some questions about how the samples taste and which ones they prefer. There is no foreseeable risk or discomfort. Participants may withdraw their data at the end of their participation if they decide that they didn't want to participate after all.

Participants' identity and personal information will be kept confidential (locked in a file cabinet to which I alone have access) and will be destroyed as soon as the study is completed. The results will be published in a scientific journal. After the study, all participants will be invited to a presentation on how taste perception works. Then each participant will be given the results of his taste test, and an opportunity to sample foods having both typical and increased amounts of the preferred tastes. We hope you will find this information useful to you in seasoning your food in the cafeteria.

Your participation in this study is strictly voluntary. You may withdraw your participation at any time. Your decision as to whether to participate will have no effect on any benefits you now receive or may need to receive in the future from any agency. For answers to questions pertaining to the research, research participants' rights, or in the event of a research-related injury, you may contact me directly, at 555-1212; Dr. John Smith, Director of Research, at 555-1313; or Dr. Mary Doe, Hospital Director, at 555-1414.

Sincerely yours,

Mary Jones, Research Psychologist

Please indicate your consent by signing a copy of this letter and returning it to me. The other copy is for you to keep.

I have read this letter and consent to participate Signature:

Date:

Procedures Involved in the Research

The Belmont Report's *principle of beneficence* states that, if possible, participation in research should directly benefit subjects and, at a minimum, do them no harm. The research procedures must be described to the IRB in detail to allow the committee to assess the potential benefits and risks entailed. Its members must be able to decide whether any risks exist, whether these risks have been minimized, and whether less risky alternative procedures could be used.

The scheduling and content of the debriefing session also should be a part of your IRB application. The IRB will want to know whether the study involves deception and, if so, how misleading aspects of the study will be explained to participants. Remember, however, that the APA code of ethics for research specifies that "psychologists never deceive research participants about

significant aspects that would affect their willingness to participate, such as physical risks, discomfort, or unpleasant emotional experiences" (APA, 1992, p. 1609).

Confidentiality

Finally in your application to the IRB describe your plan for preserving the confidentiality of the data. The *principle of beneficence* requires that participants' data and records be kept confidential to avoid risks to them, when this is the wish of the participant, or when this is guaranteed by the researcher.

Whenever possible, the data should be recorded anonymously. This usually is possible if the data from a single participant are collected all at one time. If additional data have to be collected at a future time, some means of identifying participants will have to be recorded so that the data from different sessions can be collated. Individual records can be coded by identification number; if it is necessary to record names, the list of names and identification numbers can be stored in a secure place.

Researchers have devised ingenious strategies for preserving confidentiality in sensitive research. In some studies involving inflammatory information, the list of participants has been sent to a lawyer living in a foreign country who would be able to resist subpoenas from the U.S. government, thus protecting the confidentiality of the data even from a court order (Fox & Tracy, 1986)! Most likely, your study will not be this sensitive.

APA journals require that authors keep their data for five years and, when ethically permissible, share it with other researchers who may want to replicate the research or check its data analyses. After your study is complete, therefore, it is a good idea to record your data without identification in a shareable format.

FINAL COMMENTS

With the approval of your college's IRB, you finally are ready to do the study and analyze the data to find out what happened. If you do a qualitative rather than a quantitative study (e.g., a case study, participant observation study, or phenomenological research), the analysis will involve studying the records you collect to construct themes or patterns that are consistent with the observations. The analysis here will be challenging and your creativity will be an asset.

If your study is quantitative, the analysis will be more structured than in a qualitative study; in fact, we recommend that you plan the analysis in advance of doing the study. (Some IRBs may want information on your data analysis methods to judge whether they are adequate to answer the questions your study addresses.) If you are not good at statistical analysis, it will be worth your while to sit down with an expert to outline the analysis. In the process, you may discover that you need additional measures or different treatment groups.

You should do your data analysis on a computer; it is just too easy to make numerical mistakes computing by hand or with a calculator. (Remember, you will be sharing your results, and possibly your data, with other interested scientists, so your analysis should be error free.) Your college probably has one of the major commercial statistical packages, such as SPSS or SYSTAT. If not, you can download the program *Student Statistician* from the Website for this book (see the preface).

After the data analysis is complete, your final step will be to share your

research with the scientific community. Researchers usually do this by giving presentations at scientific meetings and publishing articles in scientific journals. How this is done is the subject of the next, and final, chapter of this book.