CHAPTER 8

Tools of Research

Operational definitions transform constructs into observable measures.

INSTRUCTIONAL OBJECTIVES

After studying this chapter, the student will be able to:

1. Explain the role of measurement in research.
2. Access sources such as *Mental Measurements Yearbook* and *Tests in Print* to obtain information necessary for evaluating standardized tests and other measuring instruments.
3. State the difference between a test and a scale.
4. Distinguish between norm-referenced and criterion-referenced tests.
5. Distinguish between measures of aptitude and achievement.
6. Distinguish between ceiling effect and floor effect and discuss why these may be of concern.
7. Describe the steps to follow in preparing a Likert scale for measuring attitudes.
8. Define performance assessment and discuss its advantages and disadvantages.
9. Describe the characteristics of a bipolar adjective scale.
10. State the kinds of errors that are common to rating scales.
11. State advantages and disadvantages of self-report personality measures.
12. List at least five guidelines that a researcher should follow when using direct observation as a data-gathering technique.
13. Define a situational test, and tell when it might be used in research.
14. State the essential characteristic of a projective technique and name at least two well-known projective techniques.

One aim of quantitative research is to obtain greater understanding of relationships among variables in populations. For example, you might ask, What is the relationship between intelligence and creativity among 6-year-olds? You cannot directly observe either intelligence or creativity. Nor can you directly observe all 6-year-olds. But this does not mean that you must remain in ignorance about this and similar questions. There are observable behaviors that are accepted as being valid indicators of constructs such as intelligence and creativity. Using indicators to approximate constructs is the measurement aspect of research.

Some measurement is very straightforward, using a single indicator to represent a variable. For example, you could measure a person's educational background by asking about the highest grade he or she had completed. Similarly, such variables as grade level, nationality, marital status, or number of children could be measured by a single indicator simply because these variables refer to phenomena that are very clear and for which a single indicator provides an acceptable measure. Other variables, however, are more complex and much more difficult to measure. In these cases, using a single indicator is not appropriate.

Selecting appropriate and useful measuring instruments is critical to the success of any research study. One must select or develop scales and instruments that can measure complex constructs such as intelligence, achievement, personality, motivation, attitudes, aptitudes, interests, and self-concept. There are two basic ways to obtain these measures for your study: Use one that has already been developed or construct your own.

To select a measuring instrument, the researcher should look at the research that has been published on his or her question to determine what other researchers have used to measure the construct of interest. These reports will generally indicate whether the instrument worked well or whether other procedures might be better. Other useful sources for identifying published instruments for one's research purposes are the *Seventeenth Mental Measurements Yearbook* (Geisinger, Spies, Carlson, & Plake, 2007) and a companion volume, *Tests in Print VII* (Murphy, Plake, & Spies, 2006). Each edition of *Tests in Print* provides an index of all known commercially available tests in print at the time, with information on publisher and date of publication. A subject index helps one to locate tests in a specific category. The Buros Center for Testing website (www.unl.edu/buros) allows you to examine a large amount of information on tests and testing. Once you locate an available test, you then consult the *Mental Measurements Yearbook* for more information and a critical review of the test. The "Test Reviews Online," a service of the Buros Center for Testing, provides reviews exactly as they appear in the *Mental Measurements Yearbook* series. Another good source of information about both published and unpublished tests is the Educational Testing Service (ETS) Test Collection. The ETS Test Collection is a library of more than 20,000 commercial and research tests and other measuring devices designed to provide up-to-date test information to educational researchers. It is available on the web (www.ets.org/testcoll). ETS also has the collection *Tests in Microfiche*, which provides not only an index of unpublished tests but also copies of the tests on microfiche.

If researchers cannot find a previously developed instrument, then they must develop their own. The procedure involves identifying and using behavior that can be considered an indicator of the construct. To locate these indicators, researchers should turn first to the theory behind the construct. A good theory generally suggests how the construct will manifest itself and the changes that can be observed; that is, it suggests ways to measure the construct(s). For example, the general (g factor) theory of intelligence influenced the choice of tasks in the construction of early intelligence tests. Shavelson, Huber, and Stanton's (1976) multidimensional theory of self-concept served as the blueprint for a number of self-concept measures that have had a major influence on both theory and classroom practice. For instance, the Shavelson model was the basis for Marsh's (1988)

widely used SDQ (Self-Description Questionnaire), which measures self-concept in preadolescents, adolescents, and late adolescents/young adults. Following construction of an instrument, additional research is used to support or revise both the instrument and the theory upon which it is based. Researchers can also use their own experiences and expertise to decide on the appropriate indicators of the construct. In this chapter, we briefly discuss some of the main types of measuring instruments that are used in educational research: achievement and aptitude tests, personality tests, attitude scales, and observational techniques.

TESTS

Tests are valuable measuring instruments for educational research. A **test** is a set of stimuli presented to an individual in order to elicit responses on the basis of which a numerical score can be assigned. This score, based on a representative sample of the individual's behavior, is an indicator of the extent to which the subject has the characteristic being measured.

The utility of these scores as indicators of the construct of interest is in large part a function of the objectivity, validity, and reliability of the tests. Objectivity is the extent of agreement among scorers. Some tests, such as multiple-choice and true–false tests, are described as objective because the scoring is done by comparing students' answers with the scoring key, and scorers need make no decisions. Essay tests are less objective because scores are influenced by the judgment and opinions of the scorers. In general, validity is the extent to which a test measures what it claims to measure. Reliability is the extent to which the test measures accurately and consistently. We discuss validity and reliability in Chapter 9.

ACHIEVEMENT TESTS

Achievement tests are widely used in educational research, as well as in school systems. They are used to measure what individuals have learned. Achievement tests measure mastery and proficiency in different areas of knowledge by presenting subjects with a standard set of questions involving completion of cognitive tasks. Achievement tests are generally classified as either standardized or teacher/researcher made.

Standardized Tests

Standardized tests are published tests that have resulted from careful and skillful preparation by experts and cover broad academic objectives common to the majority of school systems. These are tests for which comparative norms have been derived, their validity and reliability established, and directions for administering and scoring prescribed. The directions are contained in the manuals provided by the test publishers. To establish the norms for these tests, their originators administer them to a relevant and representative sample. The norm group may be chosen to represent the nation as a whole or the state, city, district, or local school. The *mean* for a particular grade level in the sample becomes the norm for that grade level. It is important to distinguish between a norm and a standard. A *norm* is not necessarily a goal or a criterion of what should be. It is a

measure of what *is*. Test norms are based on the actual performance of a specified group, not on standards of performance. The skills measured are not necessarily what "ought" to be taught at any grade level, but the use of norms does give educators a basis for comparing their groups with an estimate of the mean for all children at that grade level. Currently, as part of the accountability movement, standardized tests are being widely used to measure students' achievement. The No Child Left Behind Act of 2001 mandated that states have instruments that ensure accurate measurement of a body of skills and knowledge judged to be important and that the instruments be administered and scored under standardized conditions. The measurement aims to determine the number of students at a particular grade level who know a particular set of facts or are proficient in a particular set of skills. For example, Indiana has the ISTEP (Indiana Student Test of Educational Progress), Illinois has the ISAT (Illinois Standard Achievement Test), and California has the CST (California Standards Test).

Standardized achievement tests are available for single school subjects, such as mathematics and chemistry, and also in the form of comprehensive batteries measuring several areas of achievement. An example of the latter is the California Achievement Test (CAT/5), which contains tests in the areas of reading, language, and mathematics and is appropriate for grades kindergarten to 12. Other widely used batteries include the Iowa Tests of Basic Skills (ITBS), the Metropolitan Achievement Tests (MAT-8), the SRA Achievement Series, and the Stanford Achievement Test Series (SAT-9). Some well-known single-subject achievement tests are the Gates–MacGinitie Reading Test, the Nelson–Denny Reading Test, and the Modern Math Understanding Test (MMUT). If one is interested in measuring achievement in more than one subject area, it is less expensive and time-consuming to use a battery. The main advantage of the test battery is that each subtest is normed on the same sample, which makes comparisons across subtests, both within and between individuals, easier and more accurate.

In selecting an achievement test, researchers must be careful to choose one that is reliable and is appropriate (valid) for measuring the aspect of achievement in which they are interested. There should be a direct link between the test content and the curriculum to which students have been exposed. The test must also be valid and reliable for the type of subjects included in the study. Sometimes a researcher is not able to select the test but must use what the school system has already selected. The *Mental Measurements Yearbooks* present a comprehensive listing, along with reviews of the different achievement tests available.

If an available test measures the desired behavior and if the reliability, validity, and the norms are adequate for the purpose, then there are advantages in using a standardized instrument. In addition to the time and effort saved, investigators realize an advantage from the continuity of testing procedures—the results of their studies can be compared and interpreted with respect to those of other studies using the same instrument.

Researcher-Made Tests

When using standardized tests of achievement is not deemed suitable for the specific objectives of a research study, research workers may construct their own tests. It is much better to construct your own test than to use an inappropriate

standardized one just because it is available. The advantage of a **researcher-made test** is that it can be tailored to be content specific; that is, it will match more closely the content that was covered in the classroom or in the research study. For example, suppose a teacher wants to compare the effects of two teaching methods on students' achievement in mathematics. Although there are excellent standardized tests in mathematics, they are generally designed to measure broad objectives and may not focus sufficiently on the particular skills the researcher wishes to measure. It would be wise in this case to construct the measuring instrument, paying particular attention to evidence of its validity and reliability. The researcher should administer a draft of the test to a small group who will not participate in the study but who are similar to those who will participate. An analysis of the results enables the researcher to check the test's validity and reliability and to detect any ambiguities or other problems before employing the test. For suggestions on achievement test construction, refer to specialized texts in measurement, such as those by Popham (2005), Thorndike (2005), Kubiszyn and Borich (2006), and Haladyna (2004).

Norm-Referenced and Criterion-Referenced Tests

On the basis of the type of interpretation made, standardized and **teacher-made tests** may be further classified as **norm-referenced** or **criterion-referenced.** Norm-referenced tests permit researchers to compare individuals' performance on the test to the performance of other individuals. An individual's performance is interpreted in terms of his or her relative position in a specified reference group known as the *normative group*. Typically, standardized tests are norm referenced, reporting performance in terms of percentiles, standard scores, and similar measures.

In contrast, criterion-referenced tests enable researchers to describe what a specific individual can do, without reference to the performance of others. Performance is reported in terms of the level of mastery of some well-defined content or skill domain. Typically, the level of mastery is indicated by the percentage of items answered correctly. For example, a criterion-referenced test might be used to ascertain what percentage of Indiana fourth-graders know the capitals of the 50 states. Predetermined cutoff scores may be used to interpret the individual's performance as pass–fail. The state tests used in the mandated accountability testing programs are criterion referenced. A well-known standardized instrument, the National Assessment of Educational Progress (NAEP), is criterion referenced. It is administered to a national sample of all U.S. schools to measure student knowledge in a wide variety of subject areas.

Before designing a measuring instrument, you must know which type of interpretation is to be made. In norm-referenced tests, items are selected that will yield a wide range of scores. A researcher must be concerned with the range of difficulty of the items and the power of the items to discriminate among individuals. In criterion-referenced tests, items are selected solely on the basis of how well they measure a specific set of instructional objectives. They may be easy or difficult, depending on what is being measured. The major concern is to have a representative sample of items measuring the stated objectives so that individual performance can be described directly in terms of the specific knowledge and skills that these people are able to achieve.

Test Performance Range

The range of performance that an achievement test permits is important. Researchers want a test designed so that the subjects can perform fully to their ability level without being restricted by the test. Two types of testing effects may occur. A **ceiling effect** occurs when many of the scores on a measure are at or near the maximum possible score. Tests with a ceiling effect are too easy for many of the examinees, and we do not know what their scores might have been if there had been a higher ceiling. For example, if we gave a 60-item test and most of the scores fell between 55 and 60, we would have a ceiling effect. A graph of the frequency distribution of scores would be negatively skewed (see Chapter 6).

Likewise, test performance may be restricted at the lower end of the range, resulting in a **floor effect.** A floor effect occurs when a test is too difficult and many scores are near the minimum possible score. For example, a statistics test administered as a pretest before students had a statistics class would likely show a floor effect. A graph of the frequency distribution of scores would be positively skewed. A test with a floor effect would not detect true differences in examinees' achievement either. Standardized tests typically cover a wide range of student performance, so it is not likely that many students would get all or almost all questions correct (ceiling effect) or almost all questions wrong (floor effect). A researcher should, however, consult the test manual for information about ceiling and floor effects so that he or she can select an instrument that permits a wide range of performance. Test developers gather extensive data on subjects' performance during the test standardization process. Researchers who construct their own tests can try them out with various groups and examine the results for evidence of ceiling and floor effects. If it appears that performance range is restricted, then the researcher needs to revise the test.

Performance Assessments

Another way to classify achievement tests is whether they are verbal or **performance tests.** The most common achievement tests are paper-and-pencil tests measuring cognitive objectives. This familiar format, usually administered to groups, requires individuals to compose answers or choose responses on a printed sheet. In some cases, however, a researcher may want to measure performance—what an individual can *do* rather than what he or she *knows*. Performance assessment, usually administered individually, is a popular alternative to traditional paper-and-pencil tests among educators. A performance test is a technique in which a researcher directly observes and assesses an individual's performance of a certain task and/or judges the finished product of that performance. The test taker is asked to carry out a *process* such as playing a musical instrument or tuning a car engine or to produce a *product* such as a written essay. The performance or product is judged against established criteria. An everyday example of a performance test is the behind-the-wheel examination taken when applying for a driver's license. A paper-and-pencil test covering knowledge of signs and rules for driving is not sufficient to measure driving skill. In investigating a new method of teaching science, for example, you would want to know the effect of the method not only on students' cognitive behavior but also on their learning of various laboratory procedures and techniques or their ability to complete experiments. In this case, the researcher's test would require the students to perform a real task or

use their knowledge and skills to solve a science problem. Performance assessment is important in areas such as art, music, home economics, public speaking, industrial training, and the sciences, which typically involve individuals' ability to do something or produce something. Portfolios that contain a collection of student work such as poetry, essays, sketches, musical compositions, audiotapes of speeches, and even mathematics worksheets are popular in performance assessments. They provide an opportunity for teachers and researchers to gain a more holistic view of changes in students' performance over time.

Constructing a Performance Test To create a performance test, follow these three basic steps:

1. Begin with a clear statement of the objectives and what individuals will be asked to do and the conditions under which the task will be performed. A set of test specifications listing the critical dimensions to be assessed will lead to a more comprehensive coverage of the domain. State whether there will be time limits, whether reference books will be available, and so on.
2. Provide a problem or an exercise that gives students an opportunity to perform—either a simulation or an actual task. All individuals must be asked to perform the same task.
3. Develop an instrument (checklist, rating scale, or something similar) that lists the relevant criteria to use in evaluating the performance and/or the product. Make sure that the same criteria are used for each individual's performance or product.

Performance tests are useful for measuring abilities and skills that cannot be measured by paper-and-pencil tests. However, they are time intensive and thus more expensive to administer and score.

APTITUDE TESTS

Aptitude tests differ from achievement tests in that aptitude tests attempt to measure general ability or potential for learning a body of knowledge and skills, whereas achievement tests attempt to measure the actual extent of acquired knowledge and skills in specific areas. Aptitude tests measure a subject's ability to perceive relationships, solve problems, and apply knowledge in a variety of contexts. Some critics question the distinction made between aptitude and achievement tests. They point out that an aptitude test measures achievement to some extent, and an achievement test has an aptitude element. Aptitude tests were formerly referred to as **intelligence tests**, but the latter term has declined in use because of controversy over the definition of intelligence and because people tend to associate intelligence with inherited ability. Aptitude tests should *not* be considered as measures of innate (or "pure") intelligence. As noted previously, performance on such tests partly depends on the background and schooling of the subject.

Educators have found aptitude tests useful and generally valid for the purpose of predicting school success. Many of the tests are referred to as **scholastic aptitude tests,** a term pointing out specifically that the main function of these tests is to predict school performance. Well-known aptitude tests are the ACT (American College Testing Assessment) and the SAT (Scholastic Assessment Test)

for high school students and the GRE (Graduate Record Exam) and MAT (Miller Analogies Test) for college seniors.

Researchers often use aptitude tests. Aptitude or intelligence is frequently a variable that needs to be controlled in educational experiments. To control this variable, the researcher may use the scores from a scholastic aptitude test. Of the many tests available, some have been designed for use with individuals and others for use with groups.

Individual Aptitude Tests

The most widely used individually administered instruments for measuring aptitude are the Stanford–Binet Intelligence Scale (4th ed.) and the three Wechsler tests. The Stanford–Binet currently in use is the outcome of several revisions of the device first developed in France in 1905 by Alfred Binet and Theodore Simon for identifying children who were not likely to benefit from normal classroom instruction. It was made available for use in the United States in 1916. This test originally reported an individual's mental age. Later, the concept of *intelligence quotient* (IQ) was introduced. This quotient was derived by dividing mental age (MA) by chronological age (CA) and multiplying the result by 100. The present revision of the Stanford–Binet no longer employs the MA/CA ratio for determining IQ. The IQ is found by comparing an individual's performance (score) with norms obtained from his or her age group through the use of standard scores (see Chapter 6). The latest revision of the test has 15 subtests organized into four areas: Verbal Reasoning, Quantitative Reasoning, Abstract/Visual Reasoning, and Short-Term Memory. The scores on the 15 subtests are standard scores with a mean of 50 and a standard deviation of 8. The four area scores and the total IQ score all have a mean of 100 and standard deviation of 16. The Stanford–Binet is appropriate for ages 2 years through adult.

The tests David Wechsler developed to measure aptitude now come in several forms: the Wechsler Intelligence Scale for Children—Third Edition (WISC–III, 1991), the Wechsler Adult Intelligence Scale-III (WAIS–III, 1997), and the Wechsler Preschool and Primary Scale of Intelligence-Revised (WPPSI–R, 1989), which was introduced for the 4 to 6½-year age group. The Wechsler tests yield verbal IQ scores, performance IQ scores, and full-scale IQ scores derived by averaging the verbal subtest scores, the performance subtest scores, and all subtest scores, respectively. The Wechsler scales are more popular than the Stanford–Binet primarily because they require less time to administer.

Group Tests of Aptitude

A Stanford–Binet or Wechsler test must be given by a trained psychometrician to an individual subject, a procedure expensive in both time and money. Thus, they are impractical as aptitude measures for large groups of individuals. In this situation, group tests are used. The first group test of mental ability was developed during World War I for measuring the ability of men in military service. One form of this test, the Army Alpha, was released for civilian use after the war and became the model for a number of group tests. Today, many group tests of mental aptitude are available. Among the most widely used are the Cognitive Abilities Tests (CogAT), Test of Cognitive Skills (TCS/2), and the Otis–Lennon School Ability Tests (OLSAT-7). The CogAT and the OLSAT-7 are appropriate for grades kindergarten to 12, whereas the TCS/2 is used for grades 2 to 12.

TESTING AND TECHNOLOGY

New technologies are presenting opportunities for alternatives to paper-and-pencil tests. For example, the PRAXIS I test designed to assess basic skills prior to entry into teacher education is given electronically with immediate scoring and feedback on performance provided to the examinee. A computer is also used to administer the GRE and a number of other well-known tests. Many of you may have encountered computer-based testing when you took the knowledge portion of your test to obtain a driver's license.

MEASURES OF PERSONALITY

Educational researchers often use measures of personality. There are several different types of personality measures, each reflecting a different theoretical point of view. Some reflect trait and type theories, whereas others have their origins in psychoanalytic and motivational theories. Researchers must know precisely what they wish to measure and then select the instrument, paying particular attention to the evidence of its validity. Two approaches are used to measure personality: objective personality assessment and projective personality assessment.

OBJECTIVE PERSONALITY ASSESSMENT

Self-report inventories present subjects with an extensive collection of statements describing behavior patterns and ask them to indicate whether or not each statement is characteristic of their behavior by checking *yes*, *no*, or *uncertain*. Other formats use multiple choice and true–false items. The score is computed by counting the number of responses that agree with a trait the examiner is attempting to measure. For example, someone with paranoid tendencies would be expected to answer *yes* to the statement "People are always talking behind my back" and *no* to the statement "I expect the police to be fair and reasonable." Of course, similar responses to only two items would not indicate paranoid tendencies. However, such responses to a large proportion of items could be considered an indicator of paranoia.

Some of the self-report inventories measure only one trait, such as the California F-Scale, which measures authoritarianism. Others, such as Cattell's Sixteen Personality Factor Questionnaire, measure a number of traits. Other multiple-trait inventories used in research are the Minnesota Multiphasic Personality Inventory (MMPI-2), the Guilford–Zimmerman Temperament Survey, the Mooney Problem Check List, the Edwards Personal Preference Schedule (EPPS), the Myers–Briggs Type Indicator, and the Strong Interest Inventory. A popular inventory, the Adjective Checklist, asks individuals to check from a list of adjectives those that are applicable to themselves. It is appropriate for individuals in grade 9 through adults and only takes 15 minutes to complete. It yields scores on self-confidence, self-control, needs, and other aspects of personality adjustment.

Inventories have been used in educational research to obtain trait descriptions of certain defined groups, such as underachievers and dropouts. They are useful for finding out about students' self-concepts, their concerns or problems, and their study skills and habits. Inventories have also been used in research concerned with interrelationships between personality traits and such variables as aptitude, achievement, and attitudes.

Inventories have the advantages of economy, simplicity, and objectivity. They can be administered to groups and do not require trained psychometricians. Most of the disadvantages are related to the problem of validity. The validity of self-report inventories depends in part on the respondents' being able to read and understand the items, their understanding of themselves, and especially their willingness to give frank and honest answers. As a result, the information obtained from inventories may be superficial or biased. This possibility must be taken into account when using results obtained from such instruments. Some inventories have built in validity scales to detect faking, attempts to give socially desirable responses, or reading comprehension problems.

PROJECTIVE PERSONALITY ASSESSMENT

Projective techniques are measures in which an individual is asked to respond to an ambiguous or unstructured stimulus. They are called *projective* because a person is expected to project into the stimulus his or her own needs, wants, fears, beliefs, anxieties, and experiences. On the basis of the subject's interpretation of the stimuli and his or her responses, the examiner attempts to construct a comprehensive picture of the individual's personality structure. Projective methods are used mainly by clinical psychologists for studying and diagnosing people with emotional problems. They are not frequently used in educational research because of the necessity of specialized training for administration and scoring and the expense involved in individual administration. Furthermore, many researchers question their validity primarily because of the complex scoring. The two best known projective techniques are the Rorschach Inkblot Technique and the Thematic Apperception Test (TAT). The Rorschach consists of 10 cards or plates each with either a black/white or a colored inkblot. Individuals are asked what they "see." Their responses are scored according to whether they used the whole or only a part of the inkblot or if form or color was used in structuring the response, whether movement is suggested, and other aspects. In the TAT, the respondent is shown a series of pictures varying in the extent of structure and ambiguity and asked to make up a story about each one. The stories are scored for recurrent themes, expression of needs, perceived problems, and so on. The TAT is designed for individuals age 10 years through adult. There is also a form available for younger children (Children's Apperception Test) and one for senior citizens (Senior Apperception Test).

SCALES

Scales are used to measure attitudes, values, opinions, and other characteristics that are not easily measured by tests or other measuring instruments. A **scale** is a set of categories or numeric values assigned to individuals, objects, or behaviors for the purpose of measuring variables. The process of assigning scores to those objects in order to obtain a measure of a construct is called *scaling*. Scales differ from tests in that the results of these instruments, unlike those of tests, do not indicate success or failure, strength or weakness. They measure the degree to which an individual exhibits the characteristic of interest. For example, a researcher may use a scale to measure the attitude of college students toward religion or any other topic. A number of scaling techniques have been developed throughout the years.

ATTITUDE SCALES

Attitude scales use multiple responses—usually responses to statements—and combine the responses into a single scale score. Rating scales, which we discuss later in this chapter, use judgments—made by the individual under study or by an observer—to assign scores to individuals or other objects to measure the underlying constructs.

Attitudes of individuals or groups are of interest to educational researchers. An attitude may be defined as a positive or negative affect toward a particular group, institution, concept, or social object. The measurement of attitudes presumes the ability to place individuals along a continuum of favorableness–unfavorableness toward the object.

If researchers cannot locate an existing attitude scale on their topic of interest, they must develop their own scales for measuring attitudes. We discuss two types of attitude scales: summated or Likert (pronounced *Lik'ert*) scales and bipolar adjective scales.

Likert Scales: Method of Summated Ratings

The Likert scale (1932), named for Rensis Likert who developed it, is one of the most widely used techniques to measure attitudes. A **Likert scale** (a **summated rating scale**) assesses attitudes toward a topic by presenting a set of statements about the topic and asking respondents to indicate for each whether they strongly agree, agree, are undecided, disagree, or strongly disagree. The various agree–disagree responses are assigned a numeric value, and the total scale score is found by summing the numeric responses given to each item. This total score assesses the individual's attitude toward the topic.

A Likert scale is constructed by assembling a large number of statements about an object, approximately half of which express a clearly favorable attitude and half of which are clearly unfavorable. Neutral items are not used in a Likert scale. It is important that these statements constitute a representative sample of all the possible opinions or attitudes about the object. It may be helpful to think of all the subtopics relating to the attitude object and then write items on each subtopic. To generate this diverse collection of items, the researcher may find it helpful to ask people who are commonly accepted as having knowledge about and definite attitudes toward the particular object to write a number of positive and negative statements. Editorial writings about the object are also good sources of potential statements for an attitude scale. Figure 8.1 shows items from a Likert scale designed to measure attitudes toward capital punishment.

For pilot testing, the statements, along with five response categories arranged on an agreement–disagreement continuum, are presented to a group of subjects. This group should be drawn from a population that is similar to the one in which the scale will be used. The statements should be arranged in random order so as to avoid any response set on the part of the subjects.

The subjects are directed to select the response category that best represents their reaction to each statement: *strongly agree* (SA), *agree* (A), *undecided* (U), *disagree* (D), or *strongly disagree* (SD). There has been some question regarding whether the undecided option should be included in a Likert scale. Most experts in the field recommend that the researcher include a neutral or undecided choice

1. Capital punishment serves as a deterrent to premeditated crime.
   SA A U D SD
*2. Capital punishment is morally wrong.
   SA A U D SD
3. The use of capital punishment is the best way for society to deal with hardened criminals.
   SA A U D SD
*4. I would sign a petition in favor of legislation to abolish the death penalty.
   SA A U D SD
*5. Capital punishment should not be used because there is always the possibility that an innocent person could be executed.
   SA A U D SD
6. Capital punishment reduces the use of tax monies for the care of prison inmates.
   SA A U D SD
*7. Only God has the right to take a human life.
   SA A U D SD
8. If more executions were carried out, there would be a sharp decline in violent crime.
   SA A U D SD
*9. Capital punishment should only be considered after all rehabilitation efforts have failed.
   SA A U D SD
10. I believe murder deserves a stronger penalty than life imprisonment.
   SA A U D SD
*11. Capital punishment should be abolished because it is in conflict with basic human rights.
   SA A U D SD
*12. I would be willing to participate in an all-night vigil to protest the execution of a criminal in my state.
   SA A U D SD

*These are negative items, agreement with which is considered to reflect a negative or unfavorable attitude toward capital punishment.

Figure 8.1 Example of a Likert Scale

Source: These items were taken from an attitude scale constructed by a graduate student in an educational research class.

because some respondents actually feel that way and do not want to be forced into agreeing or disagreeing.

Scoring Likert Scales To score the scale, the response categories must be weighted. For favorable or positively stated items, *strongly agree* is scored 5, *agree* is scored 4, *undecided* is scored 3, *disagree* is scored 2, and *strongly disagree* is scored 1. For unfavorable or negatively stated items, the weighting is reversed because disagreement with an unfavorable statement is psychologically equivalent to agreement with a favorable statement. Thus, for unfavorable statements, *strongly agree* would receive a weight or score of 1 and *strongly disagree* a weight of 5. (The weight values do not appear on the attitude scale presented to respondents, nor do the asterisks seen in Figure 8.1.)

The sum of the weights of all the items checked by the subject is the individual's total score. The highest possible scale score is $5 \times N$ (the number of items); the lowest possible score is $1 \times N$.

Let us consider an example of scoring a Likert scale by looking at just the first six statements of the scale shown in Figure 8.1. An individual would complete this scale by circling the appropriate letter(s) for each statement.

The following are the responses circled by a hypothetical respondent and the score for each item:

Response	Score
1. D	2
2. SA	1
3. D	2
4. A	2
5. A	2
6. U	3

The individual's total score on the six items is 12 (out of a possible 30). Divide the total score by the number of items to arrive at a mean attitude score: $2 + 1 + 2 + 2 + 2 + 3)/6 = 2.0$. Because the mean score is less than 3, we conclude that this individual has a moderately negative attitude toward capital punishment.

Item Analysis After administering the attitude scale to a preliminary group of respondents, the researcher does an **item analysis** to identify the best functioning items. The item analysis typically yields three statistics for each item: (1) an item discrimination index, (2) the percentage of respondents marking each choice to each item, and (3) the item mean and standard deviation.

The item discrimination index shows the extent to which each item discriminates among the respondents in the same way as the total score discriminates. The item discrimination index is calculated by correlating item scores with total scale scores, a procedure usually done by computer. If high scorers on an individual item have high total scores and if low scorers on this item have low total scores, then the item is discriminating in the same way as the total score. To be useful, an item should correlate at least .25 with the total score. Items that have very low correlation or negative correlation with the total score should be eliminated because they are not measuring the same thing as the total scale and hence are not contributing to the measurement of the attitude. The researcher will want to examine those items that are found to be nondiscriminating. The items may be ambiguous or double barreled (containing two beliefs or opinions in one statement), or they may be factual statements not really expressing feelings about the object. Revising these items may make them usable. The item analysis also shows the percentage of respondents choosing each of the five options and the mean and standard deviation for each item. Items on which respondents are spread out among the options are preferred. Thus, if most respondents choose only one or two of the options, the item should be rewritten or eliminated. After selecting the most useful items as indicated by the item analysis, the researcher should then try out the revised scale with a different group of subjects and again check the items for discrimination and variability.

Validity Validity concerns the extent to which the scale really measures the attitude construct of interest. It is often difficult to locate criteria to be used

in obtaining evidence for the validity of attitude scales. Some researchers have used observations of actual behavior as the criterion for the attitude being measured. This procedure is not often used because it is often difficult to determine what behavior would be the best criterion for the attitude and also because it is expensive.

One of the easiest ways to gather validity evidence is to determine the extent to which the scale is capable of discriminating between two groups whose members are known to have different attitudes (see Chapter 9). To validate a scale that measures attitudes toward organized religion, a researcher would determine if the scale discriminated between active church members and people who do not attend church or have no church affiliation. A scale measuring attitudes toward abortion should discriminate between members of pro-life groups and members of pro-choice groups. By "discriminate," we mean that the two groups would be expected to have significantly different mean scores on the scale. Another method of assessing validity is to correlate scores on the attitude scale with those obtained on another attitude scale measuring the same construct and whose validity is well established.

Reliability The reliability of the new scale must also be determined. Reliability is concerned with the extent to which the measure would yield consistent results each time it is used. The first step in ensuring reliability is to make sure that the scale is long enough—that it includes enough items to provide a representative sampling of the whole domain of opinions about the attitudinal object. Other things being equal, the size of the reliability coefficient is directly related to the length of the scale. Research shows, however, that if the items are well constructed, scales having as few as 20 to 22 items will have satisfactory reliability (often above .80). The number of items needed depends partly on how specific the attitudinal object is; the more abstract the object, the more items are needed.

You would also want to calculate an index of reliability. The best index to use for an attitude scale is coefficient alpha (see Chapter 9), which provides a measure of the extent to which all the items are positively intercorrelated and working together to measure one trait or characteristic (the attitude). Many statistical computer programs routinely calculate coefficient alpha as a measure of reliability. For further discussion on the construction of Likert and other attitude scales, the reader is referred to Mueller (1986).

Bipolar Adjective Scales

The **bipolar adjective scale** presents a respondent with a list of adjectives that have bipolar or opposite meanings. Respondents are asked to place a check mark at one of the seven points in the scale between the two opposite adjectives to indicate the degree to which the adjective represents their attitude toward an object, group, or concept. Figure 8.2 shows a bipolar adjective scale designed to measure attitude toward school. Notice that the respondent checked the extreme right position for item a and the extreme left position for item d. The adjective pairs making up a scale are listed in both directions; on some pairs the rightmost position is the most positive response, and on other pairs the leftmost position is the most positive. This is done to minimize a response set or a tendency to favor certain positions in a list of options. An individual might have a tendency to choose

School

a. bad	:	:	:	:	:	: ✓	good
b. fast	: ✓	:	:	:	:	:	slow
c. dull	:	:	:	:	: ✓	:	sharp
d. pleasant	✓ :	:	:	:	:	:	unpleasant
e. light	:	: ✓	:	:	:	:	heavy
f. passive	:	:	:	:	:	: ✓	active
g. worthless	:	:	:	:	: ✓	:	valuable
h. strong	:	:	: ✓	:	:	:	weak
i. still	:	:	:	: ✓	:	:	moving

Figure 8.2 Bipolar Adjective Scale Showing Responses of One Subject Toward the Concept "School"

the extreme right end and would check that position for each item. However, if the direction of the scale is changed in a random way so that the right end is not always the more favorable response, the individual must read each item and respond in terms of its content rather than in terms of a positional preference. The responses are scored by converting the positions checked into ratings (1 to 7). Seven represents the most positive and 1 the least positive response on each scale. The weights on each item would then be summed and averaged. In Figure 8.2, item weights are $7+6+6+7+3+7+6+4+5=51/9=5.67$. The score of 5.67 indicates a very positive attitude toward school.

The bipolar adjective scale is a very flexible approach to measuring attitudes. A researcher can use it to investigate attitudes toward any concept, person, or activity in any setting. It is much easier and less time-consuming to construct than a Likert scale. Instead of having to come up with approximately 20 statements, you need only select four to eight adjective pairs. It requires very little reading time by participants. The main difficulty is the selection of the adjectives to use. If one has a problem with this task, there are references such as Osgood, Suci, and Tannenbaum (1967) that provide lists of bipolar adjectives. It is probably better, however, to think of adjective pairs that are especially relevant to one's own project.

RATING SCALES

Rating scales present a number of statements about a behavior, an activity, or a phenomenon with an accompanying scale of categories. Observers or respondents are asked to indicate their assessment or judgment about the behavior or activity on the rating scale. For example, a teacher might be asked to rate the leadership ability of a student. The teacher would indicate his or her assessment of the student's characteristic leadership behavior by checking a point on a continuum or choosing a response category. It is assumed that raters are familiar with the behavior they are asked to assess. A numeric value may be attached to the points or categories so that an overall score could be obtained.

One of the most widely used rating scales is the **graphic scale,** in which the respondent indicates the rating by placing a check at the appropriate point on a

	Low	Medium	High
Personal appearance			
Social acceptability			
Speaking skills			

Figure 8.3 Example of a Graphic Scale

horizontal line that runs from one extreme of the behavior in question to the other. Figure 8.3 is an example of a graphic scale. The rater can check any point on the continuous line. Graphic scales usually assign numeric values to the descriptive points. Such scales are referred to as *numeric rating scales*. The speaking skills item in Figure 8.3 could look like this in a numeric scale:

1	2	3	4	5	6	7
one of the poorest speakers			an average speaker			one of the very best speakers

Category Scales

The **category scale** consists of a number of categories that are arranged in an ordered series. Five to seven categories are most frequently used. The rater picks the one that best characterizes the behavior of the person being rated. Suppose a student's abilities are being rated and one of the characteristics being rated is creativity. The following might be one category item:

How creative is this person? (check one)

exceptionally creative _____
very creative _____
not creative _____
not at all creative _____

To provide greater meaning, brief descriptive phrases are sometimes used to comprise the categories in this type of scale. Clearly defined categories contribute to the accuracy of the ratings. For example,

How creative is this person? (check one)

always has creative ideas _____
has many creative ideas _____
sometimes has creative ideas _____
rarely has creative ideas _____

Comparative Rating Scales

In using the graphic and category scales, raters make their judgments without directly comparing the person being rated to other individuals or groups. In **comparative rating scales,** in contrast, raters are instructed to make their judgment with direct reference to the positions of others with whom the individual might be compared. The positions on the rating scale are defined in terms of a given population with known characteristics. A comparative rating scale is shown in Figure 8.4. Such a scale might be used in selecting applicants for admission to graduate school. Raters are asked to judge the applicant's ability to do graduate work compared with that of all the students the rater has known. If the rating is to be valid, the judge must understand the range and distribution of abilities in the total group of graduate students.

Errors in Rating Because ratings depend on the perceptions of human observers, who are susceptible to various influences, rating scales are subject to

Area of Competency (to be rated)	Unusually low	Poorer than most students	About average among students	Better than most	Really superior	Not able to judge
1. Does this person show evidence of clear-cut and worthy professional goals?						
2. Does this person attack problems in a constructive manner?						
3. Does he or she take well-meant criticism and use it constructively?						

Figure 8.4 Example of a Comparative Rating Scale

considerable error. Among the most frequent systematic errors in rating people is the **halo effect,** which occurs when raters allow a generalized impression of the subject to influence the rating given on very specific aspects of behavior. This general impression carries over from one item in the scale to the next. For example, a teacher might rate a student who does good academic work as also being superior in intelligence, popularity, honesty, perseverance, and all other aspects of personality. Or, if you have a generally unfavorable impression of a person, you are likely to rate the person low on all aspects.

Another type of error is the **generosity error,** which refers to the tendency for raters to give subjects the benefit of any doubt. When raters are not sure, they tend to rate people favorably. In contrast, the **error of severity** is a tendency to rate all individuals too low on all characteristics. Another source of error is the **error of central tendency,** which refers to the tendency to avoid either extreme and to rate all individuals in the middle of the scale. For example, the ratings that teachers of English give their students have been found to cluster around the mean, whereas mathematics teachers' ratings of students show greater variation.

One way of reducing such errors is to train the raters thoroughly before they are asked to make ratings. They should be informed about the possibility of making these "personal bias" types of errors and how to avoid them. It is absolutely essential that raters have adequate time to observe the individual and his or her behavior before making a rating. Another way to minimize error is to make certain that the behavior to be rated and the points on the rating scale are clearly defined. The points on the scale should be described in terms of overt behaviors that can be observed, rather than in terms of behaviors that require inference on the part of the rater.

The accuracy or reliability of ratings is usually increased by having two (or more) trained raters make independent ratings of an individual. These independent ratings are pooled, or averaged, to obtain a final score. A researcher may also correlate the ratings of the two separate raters in order to obtain a coefficient of interrater reliability (see Chapter 9). The size of the coefficient indicates the extent to which the raters agree. An **interrater reliability** coefficient of .70 or higher is considered acceptable for rating scales.

DIRECT OBSERVATION

In many cases, systematic or **direct observation** of behavior is the most desirable measurement method. Observation is used in both quantitative and qualitative research. When observations are made in an attempt to obtain a comprehensive picture of a situation, and the product of those observations is notes or narratives, the research is qualitative. In Chapter 15, we discuss the use of observation in qualitative research. The current chapter focuses on observation in quantitative research where the product of using the various observational instruments is numbers. The purpose of direct observation is to determine the extent to which a particular behavior(s) is present. The observer functions like a camera or recording device to provide a record of the occurrence of the behavior in question. The researcher identifies the behavior of interest and devises a systematic procedure for identifying, categorizing, and recording the behavior in either a natural or a contrived situation. The behaviors observed in quantitative studies may be categorized as high inference and low inference. High-inference behaviors such as teacher warmth or creativity require more judgment on the part of the observer. Low-inference behaviors require less judgment by the observer. Examples of low-inference behaviors include classroom behaviors such as teachers' asking questions, praising students, or accepting students' ideas. In educational research, one of the most common uses of direct observation is in studying classroom behavior. For example, if you were interested in investigating the extent to which elementary teachers use positive reinforcement in the classroom, you could probably obtain more accurate data by actually observing classrooms rather than asking teachers about their use of reinforcement. Or if you wanted to study students' disruptive behavior in the classroom and how teachers deal with it, direct observation would provide more accurate data than reports from students or teachers.

There are five important preliminary steps to take in preparing for quantitative direct observation:

1. *Select the aspect of behavior to be observed*. Because it is not possible to collect data on everything that happens, the investigator must decide beforehand which behaviors to record and which not to record.
2. *Clearly define the behaviors falling within a chosen category*. Know what behaviors would be indicators of the attribute. In studying aggressive behavior in the classroom, would challenging the teacher or speaking out of turn be classified as aggressive, or would it be restricted to behaviors such as pushing, hitting, throwing objects, and name-calling? If observing multiple categories of behavior, make sure the categories are mutually exclusive.
3. *Develop a system for quantifying observations*. The investigator must decide on a standard method for counting the observed behaviors. For instance, establish beforehand whether an action and the reaction to it are to count as a single incident of the behavior observed or as two incidents. A suggested approach is to divide the observation period into brief time segments and to record for each period—for example, 10 seconds—whether the subject showed the behavior or not.

4. *Develop specific procedures for recording the behavior*. Record the observations immediately after they are made because observers' memory is not sufficiently reliable for accurate research. The best solution is a coding system that allows the immediate recording of what is observed, using a single letter or digit. A coding system is advantageous in terms of the observers' time and attention.
5. *Train the people who will carry out the observations*. Training and opportunity for practice are necessary so that the investigator can rely on the observers to follow an established procedure in observing and in interpreting and reporting observations. Having the observers view a videotape and discuss the results is a good training technique.

DEVICES FOR RECORDING OBSERVATIONS

Researchers use checklists, rating scales, and coding sheets to record the data collected in direct observation.

Checklists

The simplest device used is a **checklist,** which presents a list of the behaviors that are to be observed. The observer then checks whether each behavior is present or absent. A checklist differs from a scale in that the responses do not represent points on a continuum but, rather, nominal categories. For example, a researcher studying disruptive behavior would prepare a list of disruptive behaviors that might occur in a classroom. An observer would then check items such as "Passes notes to other students" or "Makes disturbing noises" each time the behavior occurs. The behaviors in a checklist should be operationally defined and readily observable.

Rating Scales

Rating scales, discussed previously, are often used by observers to indicate their evaluation of an observed behavior or activity. Typically, rating scales consist of three to five points or categories. For example, an observer studying teachers' preparation for presentation of new material in a classroom might use a scale with the following points: 5 (*extremely well prepared*), 4 (*well prepared*), 3 (*prepared*), 2 (*not well prepared*), or 1 (*totally unprepared*). A 3-point scale might include 3 (*very well prepared*), 2 (*prepared*), or 1 (*not well prepared*). Scales with more than five rating categories are not recommended because it is too difficult to accurately discriminate among the categories.

Coding Systems

Coding systems are used in observational studies to facilitate the categorizing and counting of specific, predetermined behaviors as they occur. The researcher does not just indicate whether a behavior occurred as with a checklist but, rather, uses agreed-on codes to record what actually occurred. Whereas rating scales can be completed after an observation period, coding is completed at the time the observer views the behavior.

Two kinds of coding systems are typically used by researchers: sign coding and time coding. *Sign coding* uses a set of behavior categories; each time one of the behaviors occurs, the observer codes the happening in the appropriate category. If a coding sheet used in classroom observational research listed "summarizing" as a teacher behavior, the observer would code a happening every time a teacher summarized material.

In a study using sign coding, Skinner, Buysse, and Bailey (2004) investigated how total duration and type of social play of preschool children with disabilities varied as a function of the chronological and developmental age of their social partners. They hypothesized that developmental age of each partner would better predict the duration of social play than chronological age. The 55 focal children were preschool children with mild to moderate developmental delays who were enrolled in some type of inclusive developmental day program. Each focal child was paired with 4 different same-sex partners in a standardized dyadic play situation. The observations took place outside the classroom in a specially designed and well-equipped play area. The observation consisted of two 15-minute sessions with each of the 4 play partners, or a total of 120 minutes per focal child over a period of 2 days. A video camera recorded the play behavior and trained coders used Parten's (1932) seven categories of play to code the extent to which children were engaged socially. The Battelle Developmental Inventory (Newborg, Stock, Wnek, Guidubaldi, & Svinicki, 1988) was used to assess the overall developmental status of both focal children and their social partners. A mixed-model regression analysis was employed, with the independent variables being the chronological and developmental ages of both the focal children and the partners; the dependent variable was the total duration of the category called associative play. No impact was observed for the focal children's chronological age once they accounted for developmental age. Also, they found that the influence of partner's developmental age on social play was different depending on the developmental age of the focal child. The researchers concluded that advantages accrued to preschoolers with disabilities from mixed-aged play groupings depend on the child's developmental age and those of available social partners.

In the second type of coding, called *time coding*, the observer identifies and records all predetermined behavior categories that occur during a given time period. The time period might be 10 seconds, 5 minutes, or some other period of time. Miller, Gouley, and Seifer (2004) used time coding in a study designed to document observed emotional and behavioral dysregulation in the classroom and to investigate the relationships between observed dysregulation and teachers' ratings of children's classroom adjustment and their social engagement with peers. Dysregulation was defined as emotional and behavioral displays disruptive to the preschool classroom setting. The participants were 60 low-income children attending Head Start classes. Each child was observed in a naturalistic context for two sessions of 10 minutes each, or a total of 20 minutes. The researchers used handheld computers with *The Observer* (Noldus Information Technology, 1995) software, which permitted coding of behavior along several dimensions. Analysis showed that although the majority of children did not display much dysregulated emotion or behavior in the classroom, almost one-fourth of children did display high levels of dysregulation in the observation period.

High levels of classroom dysregulation were related to teacher ratings of poor classroom adjustment and observed peer conflict behaviors, as well as negative emotional displays.

Coding has the advantage of recording observations at the time the behavior occurs, and it may yield more objective data than do rating scales. The disadvantage is that a long training period may be required for observers to learn to code behavior reliably. A number of standardized coding systems and observation forms are available. Beginning researchers should check references such as the ETS Test Collection Database for a suitable one before attempting to construct their own.

ADVANTAGES AND DISADVANTAGES OF DIRECT OBSERVATION

The most obvious advantage of systematic observation is that it provides a record of the actual behavior that occurs. We do not have to ask subjects what they would do or what they think; we have a record of their actions. Probably the most important advantage of systematic observation is its appropriateness for use with young children. It is used extensively in research on infants and on preschool children who have difficulty communicating through language and may be uncomfortable with strangers. Another advantage is that systematic observation can be used in natural settings. It is often used in educational research to study classroom or playground behavior.

The main disadvantage of systematic observation is the expense. Observations are more costly because of the time required of trained observers. Subjects may be observed for a number of sessions, requiring extended hours.

VALIDITY AND RELIABILITY OF DIRECT OBSERVATION

As with other types of measures, the validity and reliability of direct observation must be assessed. The best way to enhance validity is to carefully define the behavior to be observed and to train the people who will be making the observations. Observers must be aware of two sources of bias that affect validity: observer bias and observer effect. **Observer bias** occurs when the observer's own perceptions, beliefs, and biases influence the way he or she observes and interprets the situation. Having more than one person make independent observations helps to detect the presence of bias. **Observer effect** occurs when people being observed behave differently just because they are being observed. One-way vision screens may be used in some situations to deal with this problem. In many cases, however, after an initial reaction the subjects being observed come to pay little attention to the observer, especially one who operates unobtrusively. Some studies have used interactive television to observe classrooms unobtrusively. Videotaping for later review and coding may also be useful. Researchers who have used videotapes, for example, have found that although the children initially behaved differently with the equipment in the room, after a few days they paid no attention and its presence became routine. Handheld technologies, such as a PalmPilot, can be used to record data during observations rather than the traditional pencil-and-paper recording techniques. Professional

software such as *The Observer XT* 8.0 (Noldus Information Technology, 2008) is available for use in the collection, analysis, and presentation of observational data. Information on *The Observer XT 8.0* is available at www.noldus.com/site/doc200806003.

The accuracy or reliability of direct observation is usually investigated by having at least two observers independently observe the behavior and then determining the extent to which the observers' records agree. Reliability is enhanced by providing extensive training for the observers so that they are competent in knowing what to observe and how to record the observations. Further discussion of methods for assessing the reliability of direct observation is presented in Chapter 9.

CONTRIVED OBSERVATIONS

In **contrived observations,** the researcher arranges for the observation of subjects in simulations of real-life situations. The circumstances have been arranged so that the desired behaviors are elicited.

One form of contrived observation is the **situational test.** A classic example of a situational test—although not labeled as such at the time—was used in a series of studies by Hartshorne and May (1928) for the Character Education Inquiry (CEI). These tests were designed for use in studying the development of such behavior characteristics as honesty, self-control, truthfulness, and cooperativeness. Hartshorne and May observed children in routine school activities but also staged some situations to focus on specific behavior. For example, they gave vocabulary and reading tests to the children, collected the tests, and without the children's knowledge made duplicate copies of their answers. Later, the children were given answer keys and were asked to score their original papers. The difference between the scores the children reported and the actual scores obtained from scoring the duplicate papers provided a measure of cheating. Another test asked the children to make a mark in each of 10 small, irregularly placed circles while keeping their eyes shut. Previous control tests under conditions that prevented peeking indicated that a score of more than 13 correctly placed marks in a total of three trials was highly improbable. Thus, a score of more than 13 was recorded as evidence that the child had peeked.

Hartshorne and May (1928) found practically no relationship between cheating in different situations, such as on a test and in athletics. They concluded that children's responses were situationally specific—that is, whether students cheated depended on the specific activity, the teacher involved, and other situations rather than on some general character trait.

DATA COLLECTION IN QUALITATIVE RESEARCH

Qualitative researchers also have a number of data-gathering tools available for their investigations. The most widely used tools in qualitative research are interviews, document analysis, and observation. We discuss these methods in Chapter 15.

SUMMARY

One of the most important tasks of researchers in the behavioral sciences is the selection and/or development of dependable measuring instruments. A research study can be no better than the instruments used to collect the data. A variety of tests, scales, and inventories are available for gathering data in educational research, especially for quantitative studies. Researchers need to be aware of the strengths and limitations of these data-gathering instruments so that they can choose the one(s) most appropriate for their particular investigation. If an appropriate standardized instrument is available, the researcher would be wise to choose it because of the advantage in terms of validity, reliability, and time saved.

A test is a set of stimuli presented to an individual to elicit responses on the basis of which a numerical score can be assigned. Achievement tests measure knowledge and proficiency in a given area and are widely used in educational research. Standardized achievement tests permit the researcher to compare performance on the test to the performance of a normative reference group. Tests may be classified as paper-and-pencil or as performance tests, which measure what someone can *do* rather than what he or she *knows*. Aptitude tests are used to assess an individual's verbal and nonverbal capacities. Personality inventories are designed to measure the subject's personal characteristics and typical performance.

Attitude scales are tools for measuring individuals' beliefs, feelings, and reactions to certain objects. The major types of attitude scales are Likert-type scales and the bipolar adjective scale.

Rating scales permit observers to assign scores to the assessments made of observed behavior or activity. Among the types of rating scales are the graphic scale, the category scale, and comparative rating scales.

Rating scales, checklists, and coding systems are most commonly used to record the data in quantitative direct observation research. In coding systems, behavior can be categorized according to individual occurrences (sign coding) or number of occurrences during a specified time period (time coding).

KEY CONCEPTS

achievement test
aptitude test
attitude scale
bipolar adjective scale
category scale
ceiling effect
checklist
coding system
comparative rating scales
contrived observation
criterion-referenced test
direct observation
error of central tendency
error of severity
floor effect
generosity error
graphic scale
halo effect
intelligence test
interrater reliability
inventories
item analysis
Likert scale
norm-referenced test
observer bias
observer effect
performance test
projective technique
rating scale
researcher-made test
scale scholastic aptitude test
self-report inventories
situational test
standardized test
summated rating scale
teacher-made test
test

EXERCISES

1. What is the meaning of the term *standardized* when applied to measuring instruments?
2. What is the difference between comparative rating scales and graphic and category scales?
3. List some of the common sources of bias in rating scales.
4. What type of instrument would a researcher choose in order to obtain data about each of the following?

a. How college professors feel about the use of technology in their teaching
b. The potential of the seniors at a small college to succeed in graduate school
c. To determine if high school chemistry students can analyze an unknown chemical compound
d. How well the students at Brown Elementary School compare to the national average in reading skills
e. The advising-style preferences of a group of college freshmen
f. How well students perform in a public speaking contest
g. To determine the winner in a history essay contest
h. The general verbal and nonverbal abilities of a student with attention deficit disorder
i. The extent to which elementary teachers use negative reinforcement in the classroom, and the effect of that reinforcement on students' behavior
j. The problems faced by minority students during the first year at a large research university
k. How parents in a school system feel about moving the sixth grade from the elementary school to the middle school

5. How would you measure parents' attitudes toward a new dress code proposed for a middle school?
6. What are some procedures for increasing the accuracy of direct observation techniques?
7. Construct a five-item Likert scale for measuring peoples' attitudes toward stem cell research.
8. Intelligence tests can most accurately be described as
   a. Measures of innate mental capacity
   b. Academic achievement measures
   c. Reading tests
   d. Scholastic aptitude tests
9. List and briefly describe the instruments available for recording data in observational research.
10. What type of instrument would be most appropriate to measure each of the following?
    a. To determine if high school chemistry students can use laboratory scales to weigh specified amounts of a given chemical compound
    b. How students in the various elementary schools in Brown County compare in math skills
    c. How parents feel about an extended school day for elementary schools in the district
    d. The general verbal and nonverbal abilities of a child with dyslexia
    e. To study bullying in an elementary classroom
    f. To get a major professor's evaluation of the potential of a student for advanced work in chemistry
    g. To get a quick measure of students' attitudes toward the extracurricular programs available at the school

ANSWERS

1. *Standardized* refers to instruments for which comparative norms have been derived, their reliability and validity have been established, and directions for administration and scoring have been prescribed.
2. In judging an individual on a comparative rating scale, the rater must have knowledge of the group with which the individual is being compared. In judging an individual on graphic and category scales, raters do not make a direct comparison of the subject with other people.
3. Raters may be less than objective in judging individuals when influenced by such tendencies as the halo effect, the generosity error, the error of severity, or the error of central tendency.
4. **a.** Attitude scale
   b. Aptitude test (group)
   c. Performance test
   d. Standardized reading achievement test
   e. Inventory
   f. Rating scale (performance test)
   g. Performance test
   h. Aptitude or intelligence test (individual)
   i. Direct observation
   j. Inventory
   k. Attitude scale

5. Construct a Likert scale containing approximately 20 statements expressing positive and negative feelings about the proposed dress code or construct a bipolar adjective scale.
6. The behaviors to be observed must be specified; behaviors falling within a category must be defined; a system for quantification must be developed; and the observers must be trained to carry out the observations according to this established procedure.
7. Answers will vary.
8. d
9. Checklists indicate the presence or absence of certain behaviors. Rating scales and coding schemes both yield quantitative measures. In ratings, the person indicates his or her judgment of the behavior on a continuum. Ratings are sometimes completed in retrospect. Coding schemes are used to categorize observed behavior as it occurs.
10. **a.** Performance test
    b. Standardized achievement test
    c. Attitude scale
    d. Individual intelligence test, such as the Wechsler
    e. Observation
    f. Comparative rating scale
    g. Bipolar adjective scale

REFERENCES

Geisinger, K., Spies, R., Carlson, J., & Plake, B. (Eds.). (2007). *The Seventeenth Mental Measurements Yearbook*. Lincoln: University of Nebraska, Buros Institute of Mental Measurements.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Mahwah, NJ: Erlbaum.

Hartshorne, H., & May, M. A. (1928). *Studies in the nature of character: Studies in deceit* (Vol. 1). New York: Macmillan. [Reprinted in 1975 by Ayer, New York]

Kubiszyn, T., & Borich, G. (2006). *Educational testing and measurement*. New York: Wiley.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, no. 140.

Marsh, H. W. (1988). *Self-description questionnaire: A theoretical and empirical basis for the measurement of multiple dimensions of preadolescent self-concept: A test manual and a research monograph*. San Antonio, TX: The Psychological Corporation.

Miller, A., Gouley, K., & Seifer, R. (2004). Emotions and behaviors in the Head Start classroom: Associations among observed dysregulation, social competence, and preschool adjustment. *Early Education and Development*, *15*(2), 147–165.

Mueller, D. J. (1986). *Measuring social attitudes: A handbook for researchers and practitioners*. New York: Teachers College Press.

Murphy, L., Plake, B., & Spies, R. (Eds.). (2006). *Tests in print VII: An index to tests, test reviews, and the literature on specific tests*. Lincoln: University of Nebraska, Buros Institute of Mental Measurements.

Newborg, J., Stock, J., Wnek, L., Guidubaldi, J., & Svinicki, J. (1988). *The Battelle Developmental Inventory (BDI)*. Chicago: Riverside.

Noldus Information Technology. (1995). *The observer: System for collection and analysis of observational data* (Version 3.0). Sterling, VA: Author.

Noldus Information Technology. (2008). *The observer XT 8.0*. Sterling, VA: Author.

Osgood, C. E., Suci, G. J., & Tannenbaum, P.H. (1967). *The measurement of meaning*. Urbana: University of Illinois Press.

Parten, M. B. (1932). Social participation among preschool children. *Journal of Abnormal Social Psychology*, *27*, 243–269.

Popham, W. J. (2005). *Classroom assessment: What teachers need to know*. Boston: Allyn & Bacon.

Shavelson, R. J., Huber, J. J., & Stanton, G. C. (1976). Self-concept: Validation of construct interpretations. *Review of Educational Research*, *46*, 407–441.

Skinner, M., Buysse, V., & Bailey, D. (2004). Effects of age and developmental status of partners on play of preschoolers with disabilities. *Journal of Early Intervention*, *26*(3), 194–203.

Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education*. Upper Saddle River, NJ: Pearson Education.