CHAPTER 1

Psychological Testing and Assessment

All fields of human endeavor use measurement in some form, and each field has its own set of measuring tools and measuring units. If you're recently engaged or thinking about becoming engaged, you may have obtained an education on a unit of measure called the "carat." If you've been shopping for a computer, you may have learned something about a unit of measurement called a "byte." And if you're in need of an air conditioner, you'll no doubt want to know about the Btu (British thermal unit). Other units of measurement you may or may not be familiar with include a mile (land), a mile (nautical), a ton (long), a ton (short), a hertz, a henry, miles per hour, cycles per second, and candela per square meter. Professionals in the fields that employ these units know the potential uses, benefits, and limitations of such units in the measurements they make. So, too, users and potential users of psychological measurements need a working familiarity with the commonly used units of measurement, the theoretical underpinnings of the enterprise, and the tools employed.

Testing and Assessment

Detailed and intriguing accounts of efforts to assess people psychologically as early as the eleventh century B.C.E. in China (Yan, 1999) provide compelling testimony to the historic need for the assessment enterprise. However, the roots of contemporary psychological testing and assessment can be found in early-twentieth-century France. In 1905, Alfred Binet and a colleague published a test that was designed to help place Paris schoolchildren in appropriate classes. As history records, however, Binet's test would have consequences well beyond the Paris school district. Binet's test would serve as a catalyst to the field of psychological measurement as no test had before it. Within a decade, an English-language version of Binet's test was prepared for use in schools in the United States. In 1917, the United States declared war on Germany and entered World War I. The military needed a way to quickly screen large numbers of recruits for intellectual as well as emotional problems, and psychological testing provided the methodology. During World War II, the military would depend even more on psychological tests to screen recruits for service. The government's large-scale reliance on psychological tests served as a great impetus to the psychological testing enterprise. Following the war, an expanding number of tests purporting to measure a wide array of psychological variables burst onto the American scene.

The heyday of psychological testing was the 1950s and early 1960s. At many mental health facilities, both public and private, clients were administered groups of tests that typically included an intelligence test, a personality test, and a test to screen for neurological impairment. In the schools, the role of various psychological and educational tests in making placement and other decisions broadened. Corporate America, as well as many government agencies, also embraced psychological testing. A wide assortment of tests was being used to make critical decisions about the hiring, firing, and general utilization of personnel.

Paralleling greater reliance on data derived from psychological tests was greater public concern about such data. From the perspective of the public, psychological tests were suspect because they were so shrouded in mystery. Individuals compelled by an employer or a prospective employer to sit for a psychological test were understandably apprehensive. On the basis of data derived from the test, and for reasons not at all clear to the examinee, the testing might result in the denial of a desirable transfer or promotion, even the denial of employment. Examinees were not guaranteed any information about how well they did on the test, and they were seldom informed about the criteria on which their performance was being judged. Before long, the courts, even the Congress, would be grappling with a number of thorny questions and issues. Do psychological tests violate one's constitutional right of privacy? Do the tests really measure what they purport to measure? What kinds of decisions can and cannot be made on the basis of test data, and how should those decisions be made? What credentials, if any, are necessary to administer and interpret psychological tests? What rights do examinees undergoing psychological evaluation have?

Public scrutiny of psychological testing reached its zenith in 1965 with a series of probing and unprecedented congressional hearings (see Amrine, 1965). Against a backdrop of mounting public concern about—as well as legal challenges to—psychological testing, many psychologists in the 1960s began to look anew at the testing enterprise. Beyond being a mere instrument of measurement, a psychological test was conceptualized by many as a tool of a highly trained examiner. The value of a particular test was intimately and irrevocably linked to the expertise of the test user.

Testing and Assessment Defined

The world's receptivity to Binet's test in the early twentieth century spawned not only more tests, but more test developers, more test publishers, more test users, and the emergence of what, logically enough, has become known as a "testing" industry. "Testing" was the term used to refer to everything from the administration of a test (as in "Testing in progress") to the interpretation of a test score ("The testing indicated that . . ."). During World War I, the process of testing aptly described the group screening of thousands of military recruits. We suspect it was at that time that "testing" gained a powerful foothold in both the lay and professional vernaculars. We can find references to testing in the context of test administration and test interpretation, as well as everything in between, not only in postwar (World War I) textbooks (such as Anastasi, 1937; Bingham, 1937; Chapman, 1921; Hull, 1922; Spearman, 1927) but in varied test-related writings for decades thereafter. However, by the time of World War II, a semantic distinction between "testing" and another, more inclusive term, "assessment," began to emerge.

During World War II, the United States Office of Strategic Services (OSS) employed a variety of procedures and measurement tools—psychological tests among them—for the purpose of selecting military personnel for highly specialized positions involving spying, espionage, intelligence gathering, and the like. As summarized in *Assessment of Men* (OSS, 1948) and elsewhere (Murray & MacKinnon, 1946), the assessment data generated were subjected to thoughtful integration and evaluation by the highly trained as-

sessment center staff. The OSS model of using an innovative variety of evaluative tools, with the data derived from the evaluations analyzed by highly trained assessors, would later inspire what is now referred to as the "assessment center" approach to personnel evaluation (Bray, 1982).

Personnel evaluations, clinical evaluations, and educational evaluations are but a few of the many contexts that entail behavioral observation and active integration by an assessor of test scores and other data from various sources. In such situations, as well as other evaluations involving more than a simple test-scoring process, the term "assessment" may be preferable to "testing." Such a preference for the term "assessment" acknowledges that tests represent only one type of tool used by professional assessors. It also reflects an appreciation for the value of a test being most intimately linked with the knowledge, skill, and experience of the assessor. As Sundberg and Tyler (1962) observed, "*Tests are tools.* In the hands of a fool or an unscrupulous person they become pseudoscientific perversion" (p. 131, emphasis in the original). In many, perhaps most, evaluation contexts it is the process of assessment that breathes life and meaning into test scores; test scores are what result from testing.

Psychological Assessment, a measurement textbook by Maloney and Ward (1976), echoed the uneasiness of psychologists with the anachronistic use of "psychological testing" to describe their many varied assessment-related activities. By articulating several differences between testing and assessment, Maloney and Ward clarified the rich texture of the thoughtful, problem-solving processes of psychological assessment, which had been mistakenly clumped under the same rubric as the more technician-like tasks of psychological testing.

Maloney and Ward conceived of this problem-solving process as ever variable in nature and the result of many different factors, beginning with the reason the assessment is being undertaken. Different tools of evaluation—psychological tests among them—might be marshaled in the process of assessment depending on the particular objectives, people, and circumstances involved, as well as other variables unique to the particular situation. By contrast, psychological testing was seen as much narrower in scope, referring only to "the process of administering, scoring, and interpreting psychological tests" (Maloney & Ward, 1976, p. 9). Testing was also seen as differing from assessment because the process is "test-controlled"; decisions, predictions, or both are made solely or largely on the basis of test scores. The examiner is more key to the process of assessment, in which decisions, predictions, or both are made on the basis of many possible sources of data (including tests). Maloney and Ward also distinguished "testing" from "assessment" in regard to their respective objectives. In testing, a typical objective is to measure the magnitude of some psychological trait. For example, one might speak of "intelligence testing" if the purpose of administering a test was confined to obtaining a numerical gauge of the examinee's intelligence. In assessment, by contrast, the objective more typically extends beyond obtaining a number; rather, the aim would be to reflect the strength or absence of some psychological trait. According to this view, "assessment" would be preferable to "testing" if an evaluation of a student's intelligence was undertaken, for example, to answer a referral question about the student's ability to function in a regular classroom. Such an evaluation might explore the student's intellectual strengths and weaknesses. Further, the assessment would likely integrate the clinician's findings during the course of the intellectual evaluation that pertained to the student's social skills and judgment. Maloney and Ward (1976) further distinguished testing from assessment by noting that testing

> could take place without being directed at answering a specific referral question and even without the tester actually seeing the client or testee. For example, tests could be (and often are) administered in groups and then scored and interpreted for a variety of purposes. (p. 9)

> . . . while psychometric tests usually just add up the number of correct answers or the number of certain types of responses or performances with little if any regard for the how or mechanics of such content, clinical assessment is often far more interested in *how* the individual processes rather than the results of what he processes. The two operations, in fact, serve very different goals and purposes. (p. 39)

Regarding the collection of psychological assessment data, Maloney and Ward (1976) urged that far beyond the use of psychological tests alone, "literally, any method the examiner can use to make relevant observations is appropriate" (p. 7). Years later, Roberts and Magrab (1991) argued that assessment was not an activity to be confined to the consulting room. In presenting their community-based, interdisciplinary model for the assessment of children, they envisioned a place for traditional testing but viewed more global assessment as key to meaningful evaluation:

> Assessment in this model does not emphasize stable traits but attempts to understand a problem in the larger ecological framework in which it occurs. For assessment to be ecologically valid, a broad range of information must be collected and new methods may be required to obtain the necessary information. These methods could include routine visits to the home and the community or naturalistic observations. (p. 145)

The semantic distinction between "psychological testing" and "psychological assessment" is of more than academic interest. Society at large is best served by clear definition and differentiation between terms such as "psychological testing" and "psychological assessment," as well as related terms such as "psychological test user" and "psychological assessor." In the section "Test-User Qualifications" in Chapter 2, we argue that clear distinctions between such terms will not only serve the public good but might also help avoid the turf wars now brewing between psychology and various users of psychological tests. Admittedly, the line between what constitutes testing and what constitutes assessment is not always as straightforward as we might like it to be. However, by acknowledging that such ambiguity exists, we can work toward sharpening our definition and use of these terms; denying or ignoring their distinctiveness provides no hope of a satisfactory remedy. For our purposes, we will define **psychological assessment** as the gathering and integration of psychology-related data for the purpose of making a psychological evaluation, accomplished through the use of tools such as tests, interviews, case studies, behavioral observation, and specially designed apparatuses and measurement procedures. We will define **psychological testing** as the process of measuring psychology-related variables by means of devices or procedures designed to obtain a sample of behavior.

We elaborate on these definitions in the sections below as we discuss tests and other tools of assessment. However, having defined *assessment,* it would be useful at this juncture to define *alternate assessment.* Why? Read on.

Alternate assessment The **Individuals with Disabilities Education Act Amendments,** PL 105-17, became law in 1997. This law reauthorized and amended the Individuals with Disabilities Education Act (widely referred to as the **IDEA**), originally passed in 1975. According to Pitasky (1998), the amended IDEA "performed radical surgery on a law for which major repairs were recommended" (p. 1) and "is responsible for the most wide-sweeping and rampant changes in the history of the 27-year-old law" (p. 12). Indeed, the revisions, most of which became effective as of June 4, 1997, contained dramatic changes concerning the way that students in special education programs are educated and evaluated. Many of the provisions of the IDEA amendments are discussed elsewhere in this book. Here, let's simply point out that among other things, the new law seeks to include students with disabilities in assessments carried out at a statewide level, as well as at the

level of the individual school district. Specifically, section 612 (a) (17) of the law reads, in part, as follows:

> Children with disabilities are included in general State and district-wide assessment programs, with appropriate accommodations, where necessary. As appropriate, the State or local educational agency—(i) develops guidelines for the participation of children with disabilities in alternate assessments for those children who cannot participate in State and district-wide assessment programs; and (ii) develops and, beginning not later than July 1, 2000 conducts those alternate assessments.

The law does not expressly define "alternate assessments." However, past practice by assessors involved in evaluating students with special needs informs us what would probably pass muster with a court, should the utility of any alternate assessments be challenged. In essence, the critical question confronting assessors in special education settings may be phrased as, "What alternative assessment procedure, or adaptation of an existing procedure, shall be employed in order to assess this special education student?"

The question posed above is a familiar one to professional assessors who work in educational settings with special education students. Its answer will vary with the unique needs of each individual student. So, for example, the student who has difficulty reading the small print of a particular test may be accommodated with a large-print version of the same test or a test environment with special lighting. A student with a hearing impairment may be administered the test by means of sign language. A child with attention deficit disorder might have an extended evaluation time, with frequent breaks during periods of evaluation. So far, the process of alternate assessment may seem fairly simple and straightforward; in practice, however, it may be anything but.

Consider, for example, the case of a student with a vision impairment scheduled to be tested on a written, multiple-choice test by an alternate procedure. There are several options for the exact form of this alternate procedure. For instance, the test could be translated into Braille and administered in that form, or it could be administered by means of audiotape. Whether the test is administered by Braille or audiotape may affect the test scores—with some students doing better with a Braille administration and some doing better with an audiotaped administration. Students with superior short-term attention and memory skills for auditory stimuli would seem to have an advantage with regard to the audiotaped administration. Students with superior haptic (sense of touch) and perceptual-motor skills might have an advantage with regard to the Braille administration. We could raise a number of questions regarding the equivalence of various alternate assessments, as well as the equivalence of each of the alternate assessments to the traditional measurement method. Perhaps the key question is, "To what extent is each method really measuring the same thing?" Related questions include, "How equivalent is the alternate test to the original test?" and "How does modifying the format of a test, the time limits of a test, or any other aspect of the way a test was originally designed to be administered affect test scores?"

With this brief introduction to alternate assessment as background, we propose this definition of this somewhat elusive process: **Alternate assessment** is an evaluative or diagnostic procedure or process that varies from the usual, customary, or standardized way a measurement is derived, either by virtue of some special accommodation made to the assessee or by means of alternative methods designed to measure the same variable(s). In this definition, we have steered clear of the thorny issue of equivalence of methods; unless the alternate procedures have been thoroughly researched, there is no reason to expect that they would be equivalent—and in most cases, because the alternate procedures have been so individually tailored, there is seldom compelling research to support equivalence. State guidelines for alternate assessment will no doubt include ways of

translating measurement procedures from one format to another. Other guidelines may suggest substituting one tool of assessment, such as a test, with another tool of assessment. You might ask, "What are those other tools of assessment?"

The Tools of Psychological Assessment

The test A **test** may be defined simply as a measuring device or procedure. When the word *test* is prefaced with a modifier, what is being referred to is a measuring device or procedure designed to measure a variable related to that modifier. Consider, for example, the term *medical test,* which refers to a measuring device or procedure designed to measure some variable related to the practice of medicine (including a wide range of tools and procedures such as X rays, blood tests, and testing of reflexes). In a like manner, the term **psychological test** refers to a measuring device or procedure designed to measure variables related to psychology (for example, intelligence, personality, aptitude, interests, attitudes, and values). And whereas a medical test might involve the analysis of a sample of blood, tissue, or the like, a psychological test almost always involves the analysis of a sample of behavior. The behavior sample could range from responses to a pencil-and-paper questionnaire to oral responses to questions to performance of some task (Figure 1–1). The behavior sample could be elicited by the stimulus of the test itself or could be naturally occurring behavior (under observation).

Psychological tests may differ on a number of variables such as content, format, administration procedures, scoring and interpretation procedures, and psychometric or technical quality. The content (subject matter) of the test will, of course, vary with the focus of the particular test. But even two psychological tests purporting to measure the same construct—for example, "personality"—may differ widely in item content because of factors such as the test developer's definition of personality and the theoretical orientation of the test. For example, items on a psychoanalytically oriented personality test may have little resemblance to those on an existentially oriented personality test, yet both are "personality tests." The term **format** pertains to the form, plan, structure, arrangement, and layout of test items as well as to related considerations such as time limits. "Format" is also used to refer to the form in which a test is administered—computerized, pencil and paper, or some other form. When making specific reference to a computerized test, "format" also refers to the form of the software—IBM- or Apple-compatible. Additionally, "format" may be used with reference to the form or structure of other evaluative tools and processes, such as the conduct of interviews, the performance of tasks, and the nature of work samples and portfolios.

For sports enthusiasts, "score" typically refers to the number of points accumulated by competitors. For music aficionados, "score" refers to the written form of a musical composition. For students of psychometrics, **score** refers to a code or summary statement, usually but not necessarily numerical in nature, that reflects an evaluation with regard to performance on a test, task, interview, or some other sample of behavior. Accordingly, **scoring** is the process of assigning such evaluative codes or statements to performance on tests, tasks, interviews, or other behavior samples. As you pursue the study of the measurement of psychological and educational variables, you will learn about many different types of scores and scoring methods. You will also discover that tests differ widely in terms of their guidelines for scoring and interpretation. Some tests are designed to be scored by testtakers themselves, others are designed to be scored by trained examiners, and still others may be scored by computers. Some tests, such as most tests of intelligence, come with test manuals that are very explicit not only about scoring criteria but also about the nature of the interpretations that can be made from the

Figure 1–1
Price (and Judgment) Check in Aisle 5

Hamera and Brown (2000) described the development of a context-based Test of Grocery Shopping Skills. Designed primarily for use with persons with psychiatric disorders, this assessment tool may be useful in evaluating a skill necessary for independent living.

calculated score. Other tests, such as the Rorschach Inkblot Test (discussed in Chapter 12), are sold with no manual; the (qualified) purchaser buys the stimulus materials and then selects and uses one of many available guides for administration, scoring, and interpretation.

Tests differ with respect to their technical or psychometric quality. At this point, suffice it to say that a good test measures what it purports to measure in a consistent way and that if two tests purport to measure the exact same (identically defined) construct, the test that measures the construct better is the better (that is, the technically superior or more psychometrically sound) instrument. We have more to say about what constitutes a good test later in this chapter, and all of Part 2 is concerned with issues related to the psychometric quality of a test. Let's also note here that it is easier to identify a good test than to identify a good assessment process. A developing body of knowledge and a proving ground of experience have yielded methodologies with which tests can be evaluated for psychometric soundness. However, it is generally more difficult to evaluate the soundness of an assessment procedure because there are typically many more variables involved. Unlike a test, which may be designed to measure a particular trait, psychological assessment is undertaken in an effort to provide more information relevant to specific questions, issues, or previous conclusions, and the nature of the tools used and the procedures followed will vary accordingly. Because of the diversity of assessors' backgrounds, it is conceivable that two assessors might use entirely different sets of tools and procedures to answer any given assessment question. Can one approach to assessment be more valid than another? Yes. But determining the answer to that question with

a fair amount of certainty is sometimes an ambitious undertaking. As Maloney and Ward (1976, p. 4) put it: "We do have ways of assessing test-as-tools efficiently. On the other hand, it is much more difficult to determine the efficiency of the process of psychological assessment, primarily because there is much less agreement on what this process is or what it entails."

Consistent with common practice, we sometimes use the word "test" (as well as related terms such as "test score") in a generic sense when discussing general principles applicable to various measurement procedures. These measurement procedures range from those widely labeled as "tests" (such as paper-and-pencil examinations), to procedures that measurement experts might label with more specific terms (such as situational performance measures).

The interview Another widely used tool in the process of psychological assessment is the interview—a word that may conjure images of face-to-face talk. But an interview as a tool of psychological assessment involves more than talk. If the interview is being conducted face to face, the interviewer will probably be noting nonverbal as well as verbal behavior. For example, the interviewer may make notations regarding the interviewee's dress, manner, and eye contact. A face-to-face interview need not involve any speech if the interviewee suffers from a hearing impairment; the entire interview might be conducted in sign language. An interview may be conducted over the telephone, in which case the interviewer might make inferences regarding the content of what is said as a function of changes in the interviewee's voice quality. An interview of sorts may also be conducted by means of other electronic media, such as e-mail. In its broadest sense, then, we can define an **interview** as a method of gathering information through direct, reciprocal communication.

Interviews differ with regard to many variables, such as the purpose for which they are initiated, the time or other restrictions under which they are conducted, and the willingness of the interviewee to candidly provide information. An interview may be used by psychologists and others in clinical, counseling, forensic, or neuropsychological settings as a tool to help make diagnostic or treatment decisions. School psychologists and others in an educational setting may use interviews to help make decisions related to the appropriateness of various educational interventions or class placements. An interview may be used as a tool to help psychologists in the field of human resources to make more informed recommendations regarding the hiring, firing, and advancement of personnel. Interviews are used by psychologists who study consumer behavior to answer the questions of corporate America regarding the market for various products and services, as well as questions related to how best to advertise and promote such products and services. Researchers in psychology and related fields use interviews to explore varied psychological variables ranging from the quality of life of homeless persons (Sullivan et al., 2000), to psychological differences between Gulf War veterans with and without unexplained symptoms (Storzbach et al., 2000).

The popularity of the interview as a method for gathering information extends far beyond psychology. Just try to think of one day when you were *not* exposed to an interview on television, radio, or on the Net! However, regardless of the forum, the quality, if not the quantity of useful information produced by an interview depends to some degree on the skill of the interviewer. Interviewers differ with respect to variables such as the pacing of interviews, the extent to which they develop a rapport with interviewees, and the extent to which they convey genuineness, empathy, and a sense of humor. As you look at Figure 1–2, think about other dimensions on which you might characterize interviewers you see on television (such as juvenile versus adult, and eager-to-speak versus eager-to-listen). What types of interviewing skills do you think are necessary for

Figure 1–2
On Interviewing and Being Interviewed

Different interviewers have different styles of interviewing. How would you characterize the interview style of Howard Stern versus that of Jay Leno?

the host of a talk show? Do these skills differ from those that are necessary for a professional in the field of psychological assessment?

The portfolio In recent years, the popularity of **portfolio** (work sample) assessment in many fields, including education, has been rising. Some have argued, for example, that the best evaluation of a student's writing skills can be accomplished not by the administration of a test but by asking the student to compile a selection of writing samples. From the perspective of education administrators, portfolio assessment would seem to also have distinct advantages in assessing the effectiveness of teachers. By examining teachers' portfolios and seeing how teachers approach their coverage of various topics, educational evaluators have another tool that can help anchor judgments to work samples.

Case history data In a general sense, **case history data** refers to records, transcripts, and other accounts made in written, pictorial, or other form, in any media, that preserve archival information, official and informal accounts, as well as other data and items relevant to an assessee. Case history data may include files or excerpts from files maintained at diverse institutions and agencies such as schools, hospitals, employers, religious institutions, and criminal justice agencies. Other possible examples of case history data include letters and written correspondence, photos, family albums, newspaper or magazine clippings, and home videos, movies, and audiotapes. Work samples, artwork, doodlings, and accounts and pictures pertaining to interests and hobbies are yet other examples of case history data.

As we will see, case history data can be a very useful tool in a wide variety of assessment contexts. In a clinical evaluation, for example, case history data can be useful in shedding light relevant to an individual's past and current adjustment, as well as the events and circumstances that may have contributed to any changes in adjustment. Case history data can be of critical value in neuropsychological evaluations, where it often provides information relevant to neuropsychological functioning prior to the occurrence of a trauma or other event that results in a deficit. School psychologists rely on case history data to, among other things, answer questions about the course of a student's developmental history.

Another use of the term "case history," one synonymous with "case study," has to do with the assembly of case history data into an illustrative account. For example, a case study might detail how a number of aspects of an individual's personality combined with environmental conditions to produce a successful world leader. A case study of an individual who attempted to assassinate a high-ranking political figure might shed light on what types of individuals and conditions might lead to similar attempts in the future. In fact, as we will see in Chapter 13, The U.S. Secret Service relies heavily on behavioral case study data in its assessments of dangerousness.

Behavioral observation "To the extent that it is practically feasible, direct observation of behavior frequently proves the most clinically useful of all assessment procedures" (Goldfried & Davison, 1976, p. 44). **Behavioral observation** as a tool of assessment may be defined as monitoring the actions of others or oneself by visual or electronic means, while recording quantitative and/or qualitative information regarding the actions, typically for diagnostic or related purposes and either the design of an intervention or the measurement of the outcome of an intervention. Behavioral observation has proved to be a very useful assessment procedure, particularly in institutional settings such as schools, hospitals, prisons, and group homes. Using published or self-constructed lists of targeted behaviors, staff can observe firsthand the behavior of the person under observation and design interventions accordingly. In a school situation, for example, behavioral observation in the playground of a culturally different child suspected of having linguistic problems might reveal that the child does have English language skills but is unwilling—for reasons of shyness, cultural upbringing, or whatever—to demonstrate those abilities to an adult.

Despite the potential usefulness of behavioral observation in settings ranging from the private practitioner's consulting room to the interior of a space shuttle, it tends to be used infrequently outside institutional settings. For private practitioners, it is typically not economically feasible to spend hours out of the consulting room engaged in behavioral observation.

Role play tests Some assessment tools require assessees to role play or play themselves in some hypothetical situation and then respond accordingly. An individual being evaluated in a corporate, industrial, organizational, or military context for managerial or leadership ability, for example, might be asked to mediate a hypothetical dispute between personnel at a work site. The context of the role play may be created by various techniques ranging from live actors to computer-generated simulation. Outcome measures for such an assessment might include ratings related to various aspects of the individual's ability to resolve the conflict, such as effectiveness of approach, quality of resolution, and number of minutes to resolution.

Beyond corporate, industrial, organizational, and military settings, role play as a tool of assessment may be used in clinical settings, particularly in work with substance abusers. Clinicians may attempt to obtain a baseline measure of abuse, cravings, or cop-

ing skills by administering a **role play test** prior to therapeutic intervention, and then again at the completion of a course of treatment.

Computers as tools Traditionally, the key advantage of automated techniques has been saving assessors time in test administration, scoring, and interpretation. In interpreting test data, the ability of computers to analyze voluminous amounts of data while simultaneously comparing such data with other data in memory is especially advantageous. Related advantages of using computers in assessment include:

- *Automatic tailoring of a test's content and length for each testtaker.* Depending on their response to initial items, the content of the items testtakers are presented with may vary for each testtaker. In addition, the actual length of the test for different testtakers may also vary. The objective of this computer-adaptive testing is to tailor tests to the ability (or to the strength of some other trait) that the testtaker is presumed to possess.
- *Measurement of traits or abilities by techniques that could not be measured by more traditional methods.* For example, Mirka et al. (2000) described an assessment methodology that employs video, computer, and other components to obtain continuous assessment of back stress (Figure 1–3).
- *Quick and efficient comparisons to other testtakers.* Computers can be programmed and periodically updated with test findings for large numbers of other current or

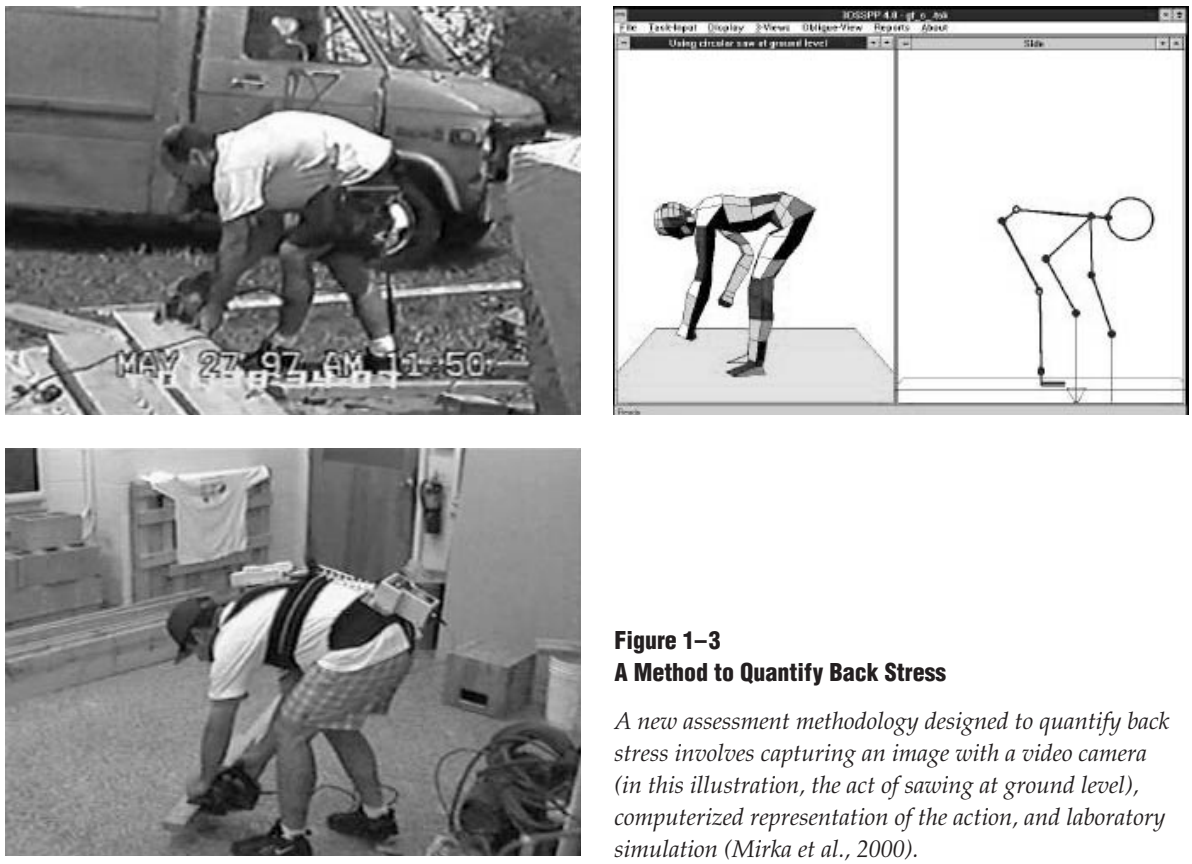

Figure 1–3
A Method to Quantify Back Stress

A new assessment methodology designed to quantify back stress involves capturing an image with a video camera (in this illustration, the act of sawing at ground level), computerized representation of the action, and laboratory simulation (Mirka et al., 2000).

previous testtakers, thus facilitating speedy comparison of results with other individuals and groups.

- *Financial savings.* In cost-conscious times, computer-assisted psychological assessment's (CAPA) promise of significant savings over time has enticed many large corporations to invest in it.

Because of the great proliferation of computerized testing, discussion of CAPA will be integrated throughout this book. In Chapter 17, we explore in detail the benefits as well as the issues that remain unresolved with regard to computer-assisted assessment.

Other tools Varied instruments of measurement can be used in psychological assessment. Video monitors wired to simple videocassette players have become more widespread as a tool of assessment. Specially created videos are used not only in job training, for example, but also in evaluating the learning and competencies of personnel. Although many math- or language-related skills can be reasonably assessed by paper-and-pencil tests, assessment by means of video adds a component of realism and attention to detail (Outtz, 1994) that is desirable in many personnel-assessment situations. Corporate managers may be asked to respond to a variety of hypothetical incidents of sexual harassment in the workplace. Police personnel may be asked about how they would respond to various types of emergencies either reenacted for the assessment video or actually recorded on tape as they happened. Psychotherapists may be asked to respond with a diagnosis and a treatment plan for each of several clients presented to them on videotape. The list of potential applications for video assessment is endless.

Psychologists and others who devise tools to assess people with disabilities and members of other special populations have been most innovative. For example, Wilson et al. (1982) described a dental plate activated by the tongue as a mechanism for test response to be used by testtakers who lack the capacity for speech or control of their hands or limbs. The device permits five kinds of response, depending on the area of the plate depressed by the tongue.

As researchers learn more about various psychology-related matters, new tools will be pressed into service to measure relevant variables. For example, a new tool in diagnosing dyslexia may be a multimedia computer device that assesses one's ability to process rapid sounds (Katz et al., 1992). Old tools may also be put to new uses based on new information. For example, ordinary blood pressure or body temperature readings may become tools of assessment in a psychological study, especially if analyzed with measures of stress or other psychological variables (see, for example, McCubbin et al., 1991; Ussher & Wilding, 1991). Biofeedback equipment is useful in obtaining measures of bodily reactions (such as muscular tension or galvanic skin response) to various sorts of stimuli. An instrument called a penile plethysmograph, which gauges male sexual arousal, has found application in sexual therapy programs with normal males experiencing sexual difficulties as well as in the treatment of sexual offenders. Impaired ability to identify odors is not uncommon in disorders such as Alzheimer's disease and Down's syndrome, in which the central nervous system (CNS) may be affected. Tests such as the University of Pennsylvania Smell Identification Test (UPSIT) have been helpful in assessing the extent of olfactory deficit in these and other diseases where there is suspected CNS involvement, such as acquired immunodeficiency syndrome (AIDS) (Brody et al., 1991). The UPSIT testtaker is sequentially exposed to 40 scratch-and-sniff odors and asked to identify each odor from a four-item word list.

There has been no shortage of innovation on the part of psychologists in devising measurement tools, or adapting existing tools, for use in psychological measurement. Yet all such tools tend to be based on a dozen or so assumptions that we now review.

Twelve Assumptions in Psychological Testing and Assessment

What follows is a listing of basic assumptions in psychological testing and assessment. Be forewarned that these assumptions are deceptively simple. One can state, for example, that psychologists who use tests to measure psychological traits assume that such traits (1) exist, (2) can be quantified, and (3) can be measured. Yet it is also true that psychologists who use tests to measure psychological traits have engaged in intense debate about the nature of the existence of psychological traits, as well as how—even if—psychological traits can be meaningfully quantified and measured. Indeed, controversy surrounds some of the most fundamental assumptions about psychological testing and assessment. As you read on, and with every successive chapter in this book, your appreciation for the complexity of the issues involved will deepen.

Assumption 1: Psychological traits and states exist. A **trait** has been defined as "any distinguishable, relatively enduring way in which one individual varies from another" (Guilford, 1959, p. 6). **States** also distinguish one person from another but are relatively less enduring (Chaplin et al., 1988).

The word *distinguishable* conveys the idea that behavior labeled with one trait term can be differentiated from behavior that is labeled with another trait term. Thus, for example, behavior within a certain context that might be viewed as religious should ideally be distinguishable from behavior within the same or another context that might be viewed as deviant. Note here that it is important to be aware of the *context* or situation in which a particular behavior is displayed when distinguishing between trait terms that may be applicable: A person who is kneeling and talking to God inside a church may be described as religious, whereas another person engaged in the exact same behavior in a public restroom might more readily be viewed as deviant. The trait term that an observer applies, as well as the strength or magnitude of the trait presumed to be present, is based on an observation of a sample of behavior. The observed sample of behavior may be obtained in a number of ways, ranging from direct observation of the assessee (such as by actually watching the individual going to church regularly and praying) to the analysis of the assessee's statements on a self-report, pencil-and-paper personality test (on which, for example, the individual may have provided an indication of great frequency in church attendance).

The phrase "relatively enduring way" in the definition serves as a reminder that a trait cannot be expected to be manifest in an individual 100% of the time. Whether a trait manifests itself, and to what degree, is presumed to depend not only on the strength of the trait in the individual but also on the nature of the situation. Stated another way, exactly how a particular trait manifests itself is, at least to some extent, situation-dependent. For example, a violent parolee may generally be prone to behave in a rather subdued way with her parole officer and much more violently in the presence of her family and friends. John may be viewed as dull and cheap by his wife but as charming and extravagant by his secretary, business associates, and others he keenly wants to impress.

The definitions of "trait" and "state" we are using also refer to a *way in which one individual varies from another.* This phrase should serve to emphasize that the attribution of a trait or state term is always a relative phenomenon. For example, in describing one person as "shy," or even in using terms such as "very shy" or "not shy," most people are typically making an unstated comparison with the degree of shyness that could reasonably be expected to be emitted by the average person under the same or similar circumstances. In psychological testing and assessment, assessors may also make such comparisons with respect to the hypothetical average person. Alternatively, assessors

may make comparisons among people who, because of their membership in some group or for any number of other reasons, are decidedly not average. As you might expect, the reference group with which comparisons are made can greatly influence one's conclusions or judgments. For example, suppose a psychologist administers a test of shyness to a 22-year-old male who earns his living as an erotic dancer. The interpretation of the test data will almost surely differ as a function of whether the reference group with which the testtaker is compared is other males in his age group or other male erotic dancers in his age group.

The term **psychological trait,** much like the term *trait* itself, covers a very wide range of possible characteristics. Thousands of psychological trait terms can be found in the English language (Allport & Odbert, 1936). Among them are psychological traits that relate to intelligence, specific intellectual abilities, cognitive style, adjustment, interests, attitudes, sexual orientation and preferences, psychopathology, personality in general, and specific personality traits. New concepts or discoveries in research may bring new trait terms to the fore. For example, a trait term seen with increasing frequency in the professional literature on human sexuality is *androgynous* (referring to a lack of primacy of male or female characteristics). Cultural evolution may bring new trait terms into common usage as it did in the 1960s when people began speaking of the degree to which women were *liberated* (or freed from the constraints of gender-dependent social expectations). A more recent example is the trait term *new age,* used in the popular culture to refer to a spiritual, almost mystical orientation.

Few people deny that psychological traits exist. Yet there has been a fair amount of controversy regarding just *how* they exist. For example, do traits have a physical existence, perhaps as a circuit in the brain? Although some have argued in favor of such a conception of psychological traits (Allport, 1937; Holt, 1971), compelling evidence to support such a view has been difficult to obtain. For our purposes, a psychological trait exists only as a **construct**—an informed, scientific idea developed or constructed to describe or explain behavior. We can't see, hear, or touch constructs, but we can infer their existence from overt behavior. In this context, "overt behavior" refers to an observable action or the product of an observable action, including test- or assessment-related responses. A challenge facing test developers is to construct tests that are at least as telling as observable behavior like that illustrated in Figure 1–4.

Assumption 2: Psychological traits and states can be quantified and measured. Amy scored 36 on a test of marital adjustment, and her husband Zeke scored 41 on the same test. *Question:* What does this information tell us about Amy, Zeke, and their adjustment to married life? *Answer:* Virtually nothing. To respond professionally to this question, we would need to know much more about (1) Amy; (2) Zeke; (3) how the construct "marital adjustment" was defined on the marital adjustment test they took; (4) the meaning of the test scores according to the test's author; and (5) research relevant to substantiating the test's guidelines for scoring and interpretation.

Test authors, much like people in general, have many different ways of looking at and defining the same phenomenon. Just think, for example, of the wide range of ways a term such as "aggressive" is used. We speak of an "aggressive salesperson," an "aggressive killer," and an "aggressive dancer," and in each of those different contexts "aggressive" carries with it a different meaning. If a personality test yields a score purporting to provide information about how aggressive a testtaker is, a first step in understanding the meaning of that score is understanding how "aggressive" was defined by the test developer. More specifically, what types of behaviors are presumed to be indicative of someone who is aggressive as defined by the test?

From a world of behaviors presumed to be indicative of the targeted trait, a test developer has a world of possible items that can be written to gauge the strength of that

Figure 1–4
Measuring Sensation Seeking

The psychological trait of sensation seeking *has been defined as "the need for varied, novel, and complex sensations and experiences and the willingness to take physical and social risks for the sake of such experiences" (Zuckerman, 1979, p. 10). A 22-item Sensation-Seeking Scale (SSS) seeks to identify people who are high or low on this trait. Assuming the SSS actually measures what it purports to measure, how would you expect a random sample of people lining up to bungee jump to score on the test, as compared with another age-matched sample of people shopping at the local mall? What are the comparative advantages of using paper-and-pencil measures, such as the SSS, and using more performance-based measures, such as the one pictured here?*

trait in testtakers.[1] For example, if the test developer deems knowledge of American history to be one component of adult intelligence, then an item that asks "Who was the second president of the United States?" may appear on the test. Similarly, if social judgment is deemed to be indicative of adult intelligence, then it would be legitimate to include an item that asks "Why should guns in the home always be inaccessible to children?" Such items having been included on an adult test of intelligence, one of the many complex issues the test developer will have to deal with is the comparative weight such items are given. Perhaps correct responses to the social judgment questions should earn more credit than correct responses to the American history questions. Perhaps, for example, a correct response to a social judgment question should be assigned a numerical value of 2 or 3 points toward the overall point total, and each correct response to the American history questions should be assigned a numerical value of 1 point. Weighting the comparative value of a test's items comes about as the result of a complex interplay among many factors, including technical considerations, the way a construct has been defined for the purposes of the test, and the value society attaches to the behaviors being evaluated.

Measurement is the assignment of numbers or symbols to characteristics of people or objects according to rules. An example of a measurement rule, this one for scoring

1. In the language of psychological testing and assessment, the word *domain* is substituted for *world* in this context. As we will see subsequently, assessment professionals speak, for example, of **domain sampling,** which may refer to either (1) a sample of behaviors from all possible behaviors that could conceivably be indicative of a particular construct, or (2) a sample of test items from all possible items that could conceivably be used to measure a particular construct.

each item on a spelling test, is "Assign the number 1 for each correct answer according to the answer key, and 0 for each incorrect answer." Another example of a measurement rule, this one for each item on a test designed to measure depression, is "Using the test's answer key as a guide, assign the number 1 for each response that indicates that the assessee is depressed, 0 for all other responses." For many varieties of psychological tests, some number representing the score on the test is derived from the examinee's responses. The test score, presumed to represent the strength of the targeted ability or trait or state, is frequently based on a cumulative model of scoring.[2] Inherent in cumulative scoring models is the assumption that the more the testtaker responds in a particular direction as keyed by the test manual as correct or consistent with a particular trait, the higher that testtaker is presumed to be on the targeted ability or trait. The rules for assigning all numbers have typically been published in the test's manual. Ideally, scientifically acceptable evidence to support the test's measurement rules, as well as all other related claims of the test author, are also included in the test's manual.

A **scale** is a set of numbers (or other symbols) whose properties model empirical properties of the objects or traits to which numbers are assigned. As we will see in Chapter 3 and again in Chapter 7, different types of scales exist, each with its own assumptions and limitations. **Scaling** may be defined as assigning numbers in accordance with empirical properties of objects or traits. Entire volumes have been written on scaling, and many different strategies of scaling can be applied in the development of a new test. An underlying assumption in all scaling efforts is that traits and abilities can be meaningfully quantified and measured. The body of professional literature on scaling provides theoretical rationales and mathematical techniques helpful in deciding how such quantification and measurement can best proceed.

Assumption 3: Various methods of measuring aspects of the same thing can be useful. A number of different tests and measurement techniques may exist to measure the same trait, state, interest, ability, attitude, or other construct, or some aspect of same. Some tests are better than others in various ways, such as the extent to which meaningful predictions can be made on the basis of the scores derived. In fact, tests can differ in a great many ways.

Tests vary in the extent to which they are linked to a theory. For example, the items for a personality test called the MMPI-2 were not developed with reference to any one theory of personality. By contrast, the items for another test, the Myers-Briggs Type Indicator, were developed on the basis of Carl Jung's theory of personality types.

Tests may also differ according to whether the items were selected on a rational or an empirical basis. As its name implies, a rational basis for a particular test item exists when the item logically taps what is known about the targeted trait. Logically, for example, we would expect people in a state of severe depression to report that they feel sad much of the time. On a rational basis, then, a test for severe depression might include a true-false item such as "I feel sad much of the time." However, test items can also be developed empirically—that is, on the basis of experience. For example, suppose researchers discovered that severely depressed people tend to agree with the following statement: "The best part of waking up is coffee in my cup." If that were the case—it is not—such a statement could be included on a strictly empirical, not rational, basis as a test item. When tests are developed empirically, the items may or may not seem to belong on the test from the standpoint of reason or logic.

There is a wide array of ways in which test items can be presented. Most familiar to you, perhaps, are items structured in a true-false, a multiple-choice, or an essay form.

2. Other, less widely used models of scoring are discussed in Chapter 7.

However, test items may be structured in other ways, so that, for example, the examinee's task is to manipulate stimulus materials by reordering or rearranging them, substituting or correcting them, or presenting them in some new form or way. A test of creative musical ability, for example, might explore the examinee's facility in manipulating a given series of musical notes.

Tests differ in their administration, scoring, and interpretation procedures. Some tests are individually administered; others are designed for group administration. Some tests have strict time limits; others are not timed. Some tests can be scored and interpreted by machines or computers; other tests are designed for submission to a committee of experts who must apply their expertise in the process of scoring and interpreting the test data.

Tests differ in the extent to which their stimulus materials are verbal or nonverbal. Tests differ in the way that they compel examinees to think and reason; success on various tests may require anything from factual recall to social judgment to great creativity—or some combination of those or other skills. Tests differ with respect to their application. One test of depression might be developed for use in an acute-care setting to identify severely depressed individuals. Another test of depression might have been developed to evaluate the effectiveness of a new drug in treating depression. In general, the utility of tests must be proved for the settings in which they were originally designed to be used, and then proved again for any additional settings in which their use is contemplated.

Assumption 4: Assessment can provide answers to some of life's most momentous questions. Every day, throughout the world, momentous questions are addressed on the basis of some type of assessment process. Is this person competent to stand trial? Who should be hired, transferred, promoted, or fired? Who should gain entry to this special program or be awarded a scholarship? Which parent shall have custody of the children? The answers to these kinds of questions are likely to have a significant impact on many lives. If they are to sleep comfortably at night, users of tests and other assessment techniques must believe that the process of assessment employed to answer such questions is fully up to the task.

Assumption 5: Assessment can pinpoint phenomena that require further attention or study. In addition to their function in evaluation for the purpose of making sometimes momentous judgments, an assumption in measurement is that tools of assessment can be used for diagnostic purposes. **Diagnosis** may be defined broadly as a description or conclusion reached on the basis of evidence and opinion through a process of distinguishing the nature of something and ruling out alternative conclusions. A **diagnostic test** may be defined as a tool used to make a diagnosis, usually for the purpose of identifying areas of deficit to be targeted for intervention.

In the field of medicine, a diagnosis is perhaps best associated with a name of some illness. A medical diagnosis may be arrived at on the basis of a physical examination, medical test data, and knowledge of the patient's medical history. In the field of education, similar tools may contribute to a comprehensive assessment. For example, a child with a reading problem may be given a thorough optometric examination and a diagnostic reading test. The resulting data will be interpreted in the context of the child's educational history. A precise statement regarding the specifics of the child's reading problem—a diagnosis—will be made.

In psychology, as in medicine, diagnosis is perhaps best associated in the public mind with the names of various illnesses, albeit mental illnesses. For example, one speaks of a diagnosis of depression or schizophrenia. In reality, however, *diagnosis* is used in a much broader sense, one that in general has to do with pinpointing psychological or behavioral phenomena, usually for further study. For example, a psychologist specializing

in measurement might use diagnostic techniques to analyze how behavior and thinking involved in taking a test administered by computer differs from behavior and thinking involved in taking that same test administered in a paper-and-pencil format. A psychologist specializing in jury research might use diagnostic techniques to analyze what is and is not compelling to a jury about various arguments. A psychologist specializing in engineering psychology might use diagnostic techniques to analyze the pros and cons of different positionings of a new control on an automobile's dashboard.

Assumption 6: Many sources of data are part of the assessment process. To understand a student, a convict, an employee, a therapy client, or any person in any role or capacity, data from a test can be helpful. However, testing and assessment professionals understand that decisions that are likely to significantly influence the course of an examinee's life are ideally made not on the basis of a single test score but, rather, from data from many different sources. Exactly what type of additional information is needed will, of course, vary with the questions the assessment procedure was initiated to answer. A partial listing of some other types of data that may be relevant to the decision-making process would include information about the examinee's current as well as past physical and mental health and academic and occupational status. Relevant family history and current family status may also make important contributions to the decision-making process, as may knowledge of the examinee's values, aspirations, and motivation.

Assumption 7: Various sources of error are part of the assessment process. In everyday conversation, we use the word *error* to refer to mistakes, miscalculations, and the like. In the context of the assessment enterprise, "error" need not refer to a deviation, an oversight, or something that otherwise violates what might have been expected. To the contrary, "error" in the context of psychological testing and assessment traditionally refers to something that is not only expected but actually considered a component of the measurement process. In this context, **error** refers to a long-standing assumption that factors other than what a test attempts to measure will influence performance on the test. Because error is a variable in any psychological assessment process, we often speak of **error variance.** Test scores earned by examinees are typically subject to questions concerning the degree to which the measurement process includes error. For example, a score on an intelligence test could be subject to debate concerning the degree to which the obtained score truly reflects the examinee's IQ, and the degree to which it was due to factors other than intelligence.

Potential sources of error are legion. An examinee's having the flu or not having the flu when taking a test is one source of error variance. In a more general sense, then, examinees are sources of error variance. Examiners, too, are sources of error variance. For example, some examiners are more professional than others in the extent to which they follow the instructions governing how and under what conditions a test should be administered. Tests themselves are another source of error variance; some tests are simply better than others in measuring what they purport to measure. There are other sources of error variance, and we will discuss them in greater detail in Chapter 5.

Instructors who teach the undergraduate measurement course will, on occasion, hear a student refer to error as "creeping into" or "contaminating" the measurement process. Yet measurement professionals tend to view error as simply an element in the process of measurement, one for which any theory of measurement must surely account. In what is referred to as "classical" or "true score" theory, an assumption is made that each testtaker has a "true" score on a test that would be obtained but for the random action of measurement error. This point is elaborated on elsewhere in this book, as well as in this chapter's *Close-up.*

CLOSE-UP

Error of Measurement and the True Score Model

Kathy applies for a job as a word processor at The Rochester Wrenchworks (TRW). To be hired, Kathy must be able to word-process accurately at the rate of 50 words per minute. The personnel office administers a total of seven brief word processing tests to Kathy over the course of seven business days. In words per minute, Kathy's scores on each of the seven tests are as follows:

52 55 39 56 35 50 54

If you were in charge of hiring at TRW and you looked at these seven scores, you might logically ask, "Which of these scores is the best measure of Kathy's 'true' word processing ability?" or, stated more succinctly, "Which is her 'true' score?"

The "true" answer to the question posed above is that we cannot say with absolute certainty from the data we have exactly what Kathy's true word processing ability is—*but* we can make an educated guess. Our educated guess would be that her true word processing ability is equal to the mean of the distribution of her word processing scores plus or minus a number of points accounted for by error in the measurement process. Error in the measurement process can be thought of as any factor entering into the process that is not directly relevant to whatever it is that is being measured. If Kathy had the misfortune on one occasion of drawing a word processor that had not been properly serviced and was of lesser quality than the other word processors she had been tested on, that is an example of error entering into the testing process. If there was excessive noise in the room on a testing occasion, if Kathy wasn't feeling well, if light bulbs blew . . . the list could go on, but the point is that any number of factors other than an individual's ability can enter into the process of measuring that ability. We can try to reduce error in a testing situation such as Kathy's by making certain, to the extent that it is possible, that all word processing equipment is functioning equally well, that the test room is free of excessive noise and has adequate lighting, and so forth. However, we can never entirely eliminate error. The best we can do is estimate how much error entered into a particular test score and then intelligently interpret the score with that information.

The tool used to estimate or infer the extent to which an observed score deviates from a true score is a statistic called the **standard error of measurement,** also known as the **standard error of a score.** In practice, few developers of tests designed for use on a widespread basis would investigate the magnitude of error with respect to a single testtaker. Typically, an average standard error of measurement is calculated for a sample of the population on which the test is designed for use. More detailed information on the nature and computation of the standard error of measurement will be presented in Chapter 5. As we will see, measures of reliability assist us in making inferences about the proportion of the total variance of test scores attributable to error variance.

Assumption 8: Tests and other measurement techniques have strengths and weaknesses. Competent test users understand a great deal about the tests they use. They understand, among other things, how a test they use was developed, the circumstances under which it is appropriate to administer the test, how the test should be administered and to whom, how the test results should be interpreted and to whom, and what the meaning of the test score is. Competent test users understand and appreciate the limitations of the tests they use, as well as how those limitations might be compensated for by data from other sources. All of this may sound quite commonsensical. It probably is. Yet this deceptively simple assumption—that test users know the tests they use and are aware of the tests' limitations—is emphasized repeatedly in the codes of ethics of associations of assessment professionals.

Assumption 9: Test-related behavior predicts non-test-related behavior. Many tests involve tasks such as blackening little grids with a number 2 pencil or simply pressing

keys on a computer keyboard. The objective of such tests typically has little to do with predicting future grid-blackening or key-pressing behavior. Rather, the objective of the test is more typically to provide some indication of other aspects of the examinee's behavior. For example, patterns of answers to true-false questions on the MMPI are used as indicators of the presence of mental disorders. The tasks in some tests mimic the actual behaviors that the test user is attempting to understand. By their nature, however, such tests yield only a sample of the behavior that can be expected to be emitted under nontest conditions. And in general, testing and assessment are conducted with the presumption that meaningful generalizations can be made from test data to behavior outside the testing situation.

Assumption 10: Present-day behavior sampling predicts future behavior. Tests sample what a person does on the day the test is administered. The obtained sample of behavior is typically used to make predictions about future behavior, such as predicted work performance of a job applicant. A rare exception to this assumption occurs in some forensic (legal) matters, where psychological tests may be used not to predict behavior but to postdict it—that is, to aid understanding of behavior that has already taken place. For example, there may be a need to understand a criminal defendant's state of mind at the time of the commission of a crime. Although it is beyond the capability of any known testing or assessment procedure to reconstruct one's state of mind, behavior samples taken at one point may be useful under certain circumstances in shedding light on the nature of one's state of mind at some point in the past. Additionally, other tools of assessment, such as case history data or the defendant's personal diary during the period in question, might all be of great value in such an evaluation.

Assumption 11: Testing and assessment can be conducted in a fair and unbiased manner. If we had to pick the one of these 12 assumptions that is more controversial than the remaining 11, this one is it. Decades of court challenges to various tests and testing programs have sensitized test developers and users to the societal demand that tests be developed so as to be fair and that tests be used in a fair manner. Today, all major test publishers strive to develop instruments that, when used in strict accordance with guidelines in the test manual, are fair. One source of fairness-related problems is the test user who attempts to use a particular test with people whose background and experience are different from the background and experience of people for whom the test was intended. In such instances, it is useful to emphasize that tests are tools that, like other, more familiar tools (hammers, ice picks, shovels, and so on), can be used properly or abused.

Some potential problems related to test fairness are more political than psychometric in nature, such as the use of tests in various social programs. For example, heated debate often surrounds affirmative action programs in selection, hiring, and access or denial of access to various opportunities. In many cases, the real question to be debated is, "What do we as a society wish to accomplish?" not "Is this test fair?"

Assumption 12: Testing and assessment benefit society. At first glance, the prospect of a world devoid of testing and assessment might seem very appealing, especially from the perspective of a harried student preparing for a week of midterm examinations. Yet a world without tests would most likely turn out to be more of a nightmare than a dream. In such a world, people could hold themselves out to the public as surgeons, bridge builders, or airline pilots regardless of their background, ability, or professional credentials. In a world without tests, teachers and school administrators could arbitrarily place children in different types of special classes simply because that is where they believed the children belonged. Considering the many critical decisions that are based on testing

and assessment procedures, as well as the possible alternatives (including decision making on the basis of human judgment, nepotism, and the like), we can readily appreciate the need for the assessment enterprise and be thankful for its existence.

Who, What, and Why?

Who are the parties in the assessment enterprise? What types of settings are assessments conducted in? Why is assessment conducted? Think about the answer to each of these important questions before reading on. Then, check your own ideas against those that follow.

Who Are the Parties?

The primary parties to the assessment enterprise are developers and publishers of tests or other methods of assessment, users of tests and other methods of assessment, and people who are evaluated by means of tests and other methods of assessment. A fourth and frequently overlooked party is society at large. Referring to these parties, respectively, as (1) the test developer, (2) the test user, (3) the testtaker, and (4) society at large, let's take a closer look at each in the context of the assessment enterprise.

The test developer Test developers create tests or other types of methods of assessment. The **American Psychological Association (APA)** estimates that upward of 20,000 new psychological tests are developed each year (APA, 1993). Among these new tests are some that were created for a specific research study, some that were created in the hope that they would be published, and some that represent refinements or modifications of existing tests.

The people who create tests bring a wide array of backgrounds, skills, and interests to the test development process. To learn more about them and the test development process, we sent letters requesting biographical information to a number of developers of some famous and not so famous tests. We inquired about major influences on these people, noteworthy aspects of the test development process, and the pros and cons of being a test developer. Many of these profiles provide not only a fascinating biographical sketch but an intriguing inside look at the test development process. Space limitations precluded us from presenting the profiles here. However, *Test Developer Profiles* can be accessed at our Internet Web site: *www.mhhe.com/psychtesting*.

Recognizing that tests and the decisions made as a result of their administration can have a significant impact on testtakers' lives, a number of professional organizations have published standards of ethical behavior that specifically address aspects of responsible test development and use. Perhaps the most detailed document addressing such issues is one jointly written by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (NCME). Referred to by many psychologists simply as "the *Standards*," *Standards for Educational and Psychological Testing* covers issues related to test construction and evaluation, test administration and use, and special applications of tests, such as special considerations when testing linguistic minorities. The *Standards* is an indispensable reference work for professional users and developers of psychological and educational tests. Initially published in 1954, revisions of the *Standards* were published in 1966, 1974, 1985, and 1999.

The test user Tests are used by a wide range of professionals, including clinicians, counselors, human resources personnel, and teachers and other school personnel. The *Standards,* as well as the official guidelines of various other professional organizations, have much to impart to test users about how, why, and the conditions under which tests should be used. For example, the principles of professional ethics promulgated by the National Association of School Psychologists (Jacob-Timm & Hartshorne, 1998) stress that school psychologists should select and use the test or tests that are most appropriate for each individual student. NASP (2000) further emphasizes that any questions that serve to prompt the psychological assessment of students be answered in as comprehensive a manner as possible—that is, with as much background information and other data as possible, including data from behavioral observation.

The test user has ethical obligations that must be fulfilled even before any testtaker is exposed to a test. For example, the test must be stored in a way that reasonably ensures that its specific contents will not be made known in advance—leaving open the possibilities of irregularities later. Note that we used the term *specific contents* in describing what must be secured from testtakers in advance of the test. In the case of some specific types of tests, mostly tests of achievement, acquainting the testtaker with the general type of questions the test will contain helps to lift the veil of mystery that may surround a test and minimize the associated test anxiety (see, for example, the booklets prepared for prospective Scholastic Aptitude Test or Graduate Record Examination examinees). With some types of tests, such as intelligence tests and projective tests of personality, such pretest descriptions of the test materials would not be advisable because they might compromise the resulting data. Another obligation of the test user before the test's administration is to ensure that a prepared and suitably trained person administers the test properly. The test administrator (or examiner) must be familiar with the test materials and procedures and have at the test site all the materials needed to properly administer the test—a sufficient supply of test protocols and other supplies, a stopwatch, if necessary, and so forth.[3] The test examiner must also ensure that the room in which the test will be conducted is suitable and conducive to the testing (Figure 1–5). To the extent that it is possible, distracting conditions such as excessive noise, heat, cold, interruptions, glaring sunlight, crowding, inadequate ventilation, and so forth should be avoided. Even a badly scratched or graffiti-grooved writing surface on a desk can act as a contaminating influence on the test administration; if the writing surface is not reasonably smooth, the written productions made on it may in some instances lead a test scorer to suspect that the examinee had a tremor or some type of perceptual-motor deficit. In short, if the test is a standardized one, it is the obligation of the test administrator to see that reasonable testing conditions prevail during the test administration; if for any reason those conditions did not prevail during an administration of the test (for instance, there was a fire drill or a real fire), an accounting of such unusual conditions should be enclosed with the test record.

Especially in one-on-one or small-group testing, rapport between the examiner and examinee is important. In the context of the testing situation, **rapport** may be defined as a working relationship between the examiner and the examinee. Such a working relationship can sometimes be achieved with a few words of small talk when examiner and examinee are introduced. If appropriate, some words regarding the nature of the test as well as why it is important for examinees to do their best may also be helpful. In other

3. **Protocol** in everyday usage refers to diplomatic etiquette. A less common usage of the word is as a synonym for the first copy or rough draft of a treaty or other official document before its ratification. This second meaning comes closer to the way the word is used with reference to psychological tests, as a noun referring to the form or sheet on which the testtaker's responses have been entered.

Figure 1–5
Less-Than-Optimal Testing Conditions

In 1917, new Army recruits sat on the floor as they were administered the first group tests of intelligence—not ideal testing conditions by current standards.

instances, as with the case of a frightened child, the achievement of rapport might involve more elaborate techniques such as engaging the child in play or some other activity until the child is deemed to have acclimated to the examiner and the surroundings. It is important that attempts to establish rapport with the testtaker not compromise any rules of the test's standardized administration instructions.

Evidence exists to support the view that, depending on the test, examiners themselves may have an effect on test results. Whether the examiner is familiar or a stranger (Sacks, 1952; Tsudzuki et al., 1957), whether the examiner is present or absent (Bernstein, 1956), and the general manner of the examiner (Exner, 1966; Masling, 1959; Wickes, 1956) are some factors that may influence performance on ability as well as personality tests (see also Cohen, 1965; Kirchner, 1966; Masling, 1960). In assessing children's abilities, the effect of examiner sex, race, and experience has been studied with a mixed pattern of results (Lutey & Copeland, 1982). Whereas some studies have indicated that students receive higher scores from female than from male examiners (for example, Back & Dana, 1977; Gillingham, 1970; Samuel, 1977), others have found that the key variable is whether the examiner and student are of the same or opposite sex. For example, Smith, May, and Lebovitz (1966) and Cieutat (1965) found that students perform better with examiners of the opposite sex, but Pedersen, Shinedling, and Johnson (1968) found that students perform better with examiners of the same sex. Examiner race and experience have been

examined in a number of studies, and reviews of these studies have concluded that these variables have little effect on student performance (Sattler, 1988; Sattler & Gwynne, 1982).

No matter how psychometrically sound a test is, the purpose of the test will be defeated if the test user fails to competently manage all phases of the testing or assessment process. For that reason alone, it is undeniably necessary for all test users, as well as all potential users of tests, to have a working familiarity with principles of measurement.

The testtaker Testtakers approach an assessment situation in different ways, and test users must be sensitive to the diversity of possible responses to a testing situation. On the day of test administration, testtakers may vary on a continuum with respect to numerous variables, including:

- The amount of test anxiety they are experiencing and the degree to which that test anxiety might significantly affect the test results.
- Their capacity and willingness to cooperate with the examiner or to comprehend written test instructions.
- The amount of physical pain or emotional distress being experienced.
- The amount of physical discomfort brought on by not having had enough to eat, having had too much to eat, or other physical conditions.
- The extent to which they are alert and wide awake as opposed to nodding off.
- The extent to which they are predisposed to agreeing or disagreeing when presented with stimulus statements.
- The extent to which they have received prior coaching.
- The importance they may attribute to portraying themselves in a good—or bad—light.
- The extent to which they are, for lack of a better term, "lucky" and can "beat the odds" on a multiple-choice achievement test (even though they may not have learned the subject matter).

As we will see, testtakers have a number of rights in assessment situations. For example, testtakers have the right to informed consent to testing, the right to have the results of the testing held confidential, and the right to be informed of the findings.

Before leaving the subject of "testtaker" as a party in the assessment process, let us make brief mention of the very rare and exceptional case where the person being assessed is deceased. Such is the case in what has been referred to as a **psychological autopsy,** or a reconstruction of a deceased individual's psychological profile on the basis of archival records, artifacts, and interviews previously conducted with the assessee or people who knew the assessee. For interested readers, a fascinating case study that employed the technique of psychological autopsy is presented by Neagoe (2000).

Society at large

> The uniqueness of individuals is one of the most fundamental characteristic facts of life. . . . At all periods of human history men have observed and described differences between individuals. . . . But educators, politicians, and administrators have felt a need for some way of organizing or systematizing the many-faceted complexity of individual differences. (Tyler, 1965, p. 3)

The societal need for "organizing" and "systematizing" has historically manifested itself in such varied questions as "Who is a witch?" "Who is schizophrenic?" and "Who is qualified?" The nature of the specific questions asked has shifted with societal concerns. The methods used to determine the answers have varied throughout history as a

function of factors such as intellectual sophistication and religious preoccupation. Palmistry, podoscopy, astrology, and phrenology, among other pursuits, have had proponents who argued that the best means of understanding and predicting human behavior was through the study of the palms, the feet, the stars, bumps on the head, tea leaves, and so on. Unlike such pursuits, the assessment enterprise has roots in science. Through systematic and replicable means that can produce compelling evidence, the assessment enterprise responds to what Tyler (1965, p. 3) referred to as the societal "need for some way of organizing or systematizing the many-faceted complexity of individual differences."

Other parties Beyond the four primary parties we have focused on here, let's briefly make note of others who may participate in varied ways in the testing and assessment enterprise. Organizations, companies, and governmental agencies sponsor the development of tests for various reasons, such as to certify personnel. Companies and services offer test scoring or interpretation services. In some cases, these companies and services are simply extensions of test publishers, and in other cases they are independent. There are people whose sole responsibility has to do with the marketing and sales of tests. Sometimes these people are employed by the test publisher, sometimes they are not. There are academicians who review tests and make evaluations as to their psychometric soundness. All of these people may also be considered parties to the enterprise.

In What Types of Settings Are Assessments Conducted and Why?

Educational settings From your own experience, you are probably no stranger to the many types of tests administered in the classroom. You have taken achievement tests—some constructed by teachers, others constructed by measurement professionals. You may have taken tests designed to assess your ability, aptitude, or interest with respect to a particular occupation or course of study. You may have also taken a group-administered test of intelligence, now also termed a **school ability test.** Such tests are frequently administered, in part to help identify children who may not be achieving at a level commensurate with their capability. Where appropriate, further evaluation with more specialized instruments may follow to assess the need for special education intervention. *Public Law 94-142* mandated that appropriate educational programs be made available to individuals with disabilities between the ages of 3 and 21 who require special education. *Public Law 99-457* specified that services be delivered to preschoolers with disabilities (birth to age 2) and encouraged services to at-risk infants, toddlers, and their families.

Tests are often used in educational settings to diagnose learning or behavior problems or both and to establish eligibility for special education programs. Individually administered intelligence and achievement measures are most often used for diagnostic purposes and are generally administered by school psychologists, psychoeducational diagnosticians, or similarly trained professionals. Interviews, behavioral observation, self-report scales, and behavior checklists are also widely used in educational settings.

Another variety of assessment that takes place daily throughout the country, in every classroom, and at every educational level is informal assessment. Evidence of such assessment comes not in test scores but in a variety of ways ranging from a sincere, enthusiastic "Good!" verbalized by instructors to nonverbal expressions of disappointment. As complex and interesting as the study of informal assessment may be, this text will limit its scope to testing and assessment of the more formal variety.

In recent years, we have witnessed the birth of a new type of achievement test: a certification of education. Particularly at the high school level, students in some areas of the country are being evaluated at the end of their course of study to determine if they indeed have acquired the minimal knowledge and skills expected of a high school

graduate. Students unable to pass this certification test receive a certificate of attendance as opposed to a high school diploma. Needless to say, the cutting score (in this case, the dividing line between passing and failing) on such a test is one with momentous consequences, and its determination must be made only by persons with a very sound technical knowledge of tests and measurement.

Another type of test administered in educational settings is that used for educational selection. Many colleges and universities require scores on standardized tests such as the Scholastic Aptitude Test (SAT) or the Graduate Record Examination (GRE) as part of the undergraduate or graduate school admission process. Foreign applicants to North American universities may be required to take a standardized test of English proficiency as part of their admission application. Few, if any, universities rely solely on standardized test scores in making admissions decisions. Typically, such decisions are based on an assessment of a number of factors ranging from grade-point average to letters of recommendation to written statements by the applicant to extracurricular interests and activities. To fulfill affirmative action requirements, variables such as ethnic background and gender may sometimes enter into the admission decision as well. Chapter 10 covers in detail psychological testing and assessment in educational settings.

Counseling settings The use of assessment in a counseling context may occur in environments as diverse as schools, prisons, or government or privately owned institutions. Regardless of where it is done, assessment is typically undertaken to identify various strengths or weaknesses, with the ultimate objective being an improvement in the assessee's adjustment, productivity, and general quality of life. Measures of social and academic skills or abilities and measures of personality, interest, attitudes, and values are among the many types of tests that a counselor might administer to a client. Objectives in testing for counseling purposes vary with stage of life and particular situation; questions to be answered range from "How can this child work and play better with other children?" to "What career is the client best suited for?" to "What activities are recommended for retirement?" Because the testtaker is in many instances the primary recipient and user of the data from a test administered by a counselor, it is imperative that a well-trained counselor fully explain the test results. Alternatively, the results of the test should be readily interpretable by testtakers themselves through easy-to-follow instructions.

Clinical settings Tests and other methods of assessment (such as interviews, case studies, and behavioral observation) are widely used in clinical settings such as inpatient and outpatient clinics; public, private, and military hospitals; private-practice consulting rooms; schools; and other institutions to screen for or diagnose behavior problems. Situations that might call for tests and other tools of clinical assessment include the following:

- A private psychotherapy client wishes to be evaluated to see if the assessment can provide any nonobvious clues regarding his maladjustment.
- A school psychologist clinically evaluates a child experiencing learning difficulties to determine if her problem lies in a deficit of ability, a problem of adjustment, a discrepancy between teaching techniques being employed and the child's favored receptive and expressive modalities, or some combination of such factors.
- A psychotherapy researcher uses assessment procedures to determine if a particular method of psychotherapy is effective in treating a particular problem.
- A psychologist-consultant retained by an insurance company is called on to give an opinion as to the reality of a client's psychological problems; is the client really experiencing such problems or malingering?

- A court-appointed psychologist is asked to give an opinion as to a defendant's competency to stand trial.
- A prison psychologist is called on to give an opinion regarding the extent of a convicted violent prisoner's rehabilitation.

The tests employed in clinical settings may be intelligence tests, personality tests, neuropsychological tests, or other specialized instruments, depending on the presenting or suspected problem area. The hallmark of testing in clinical settings is that the test or measurement technique is employed with only one individual at a time; group testing can be used only for screening at best—identifying those individuals who require further diagnostic evaluation. In Chapter 13 and elsewhere, we will look at the nature, uses, and benefits of clinical assessment.

Business settings In the business world, tests are used in many areas, particularly human resource management. As we will see in Chapter 16, personnel psychologists use tests and measurement procedures to assess whatever knowledge or skills an employer needs to have assessed—be it the ability of a prospective air traffic controller to sustain attention to detail for hours on end or the ability of a prospective military officer to lead others. A wide range of achievement, aptitude, interest, motivational, and other tests may be employed in the decision to hire as well as in related decisions regarding promotions, transfer, performance or job satisfaction, and eligibility for further training. Engineering psychologists also employ a variety of existing and specially devised tests to help people at home and in the workplace, in part by designing ergonomically efficient consumer and industrial products—products ranging from office furniture to spaceship cockpit layout.[4]

Another example of a business-related application of testing and assessment is in the area of consumer psychology. Consumer psychologists help corporate America in the development, marketing, and sale of products. Using tests as well as other techniques, psychologists who specialize in this area may be involved in taking the pulse of consumers—helping to predict the public's receptivity to a new product, a new brand, or a new advertising or marketing campaign. "What type of advertising will appeal to which type of individual?" Tests of attitudes and values have proved to be one valuable source of information to consumer psychologists and marketing professionals who endeavor to answer such questions.

Other settings Testing and assessment procedures are used in many other areas. Credentialing professionals is one such area. Before they are legally entitled to practice medicine, physicians must pass an examination. Law school graduates cannot hold themselves out to the public as attorneys until they pass their state's bar examination. Psychologists, too, must pass an examination entitling them to present themselves to the public as psychologists. And just as physicians can take further training and a test indicating that they are "Board certified" in a particular area, so can psychologists specializing in

4. "Ergonomically efficient"? An **erg** is a unit of work and **ergonomics** is the study of work; more specifically in the present context, it is the relationship between people and tools of work. Among other endeavors, engineering psychologists are involved in designing things so that we can see, hear, reach, or generally use them better. For example, it was through extensive research by engineering psychologists that the division of letters and numbers that appears on a telephone was derived. Interested in obtaining a firsthand look at the kind of work engineering psychologists do? Take a moment to look through journals like *Ergonomics, Applied Ergonomics,* and *Man-Environment Systems* next time you're in your university library.

certain areas be evaluated for a diploma from the American Board of Professional Psychology (**ABPP**) to recognize excellence in the practice of psychology. Another organization, the American Board of Assessment Psychology (**ABAP**), awards its diplomate to test users, test developers, and others who have distinguished themselves in the field of testing and assessment.

Measurement may play an important part in program evaluation—be it a large-scale government program or a small-scale privately funded one. Is the program working? How can the program be improved? Are funds being spent in the areas where they ought to be spent? These are the types of general questions that tests and measurement procedures used in program evaluation are designed to answer.

Psychological assessment plays a valuable role in the process of psychological theory building; tests and measures may be employed in basic research to confirm or disprove hypotheses derived from behavioral theories. Tests, interviews, and other tools of assessment may be used to learn more about the organization of psychological traits and serve as vehicles by which new traits can be identified.

The courts rely on psychological test data and related expert testimony as one source of information to help answer important questions such as "Is this convict competent to be executed?" "Is this parent competent to take custody of the child?" and "Did this defendant know right from wrong at the time the criminal act was committed?" Issues such as these are covered in the forensic psychology section of Chapter 13.

Issues about testing people with disabling conditions have become increasingly prominent in recent years, and our survey of these issues as well as a glimpse at specialized measurement procedures used in this area appears in Chapter 15. In Chapter 14 we detail some of the methods used by neuropsychologists to help in the diagnosis and treatment of neuropsychological deficits.

In addition to bringing you a firsthand look at the test development process, we also want to provide a glimpse of test use "in the trenches." We sent out letters to colleagues who use tests, requesting a paragraph or two about how and why they use them. Interested readers will find these responses at our *Test User Forum* on the Internet at *www.mhhe.com/psychtesting*. And by the way, if you are a user of psychological or educational tests and would like your essay posted on that site, please write to us care of our publisher.

Evaluating the Quality of Tests

We know which psychological tests are most frequently used (Archer et al., 1991; Hutton et al., 1992; Lees-Haley et al., 1996; Lubin et al., 1985; Piotrowski & Keller, 1989, 1992; Piotrowski & Lubin, 1990; Sweeney et al., 1987), but we need to know which tests are good. This of course raises a question.

What Is a Good Test?

Purely from a logical standpoint, the criteria for a good test would include clear instructions for administration, scoring, and interpretation. It would also seem to be a plus if a test offered economy in the time it takes to administer, score, and interpret it. Most of all, a good test would seem to be one that measures what it purports to measure. Ideally, the results of the assessment procedure lead to an improved quality of life for the testtaker and others.

Beyond simple logic, there are technical criteria that assessment professionals use to evaluate the quality of tests and other measurement procedures. These technical

considerations have to do with psychometrics. Synonymous with *psychometry,* **psychometrics** may be defined as the science of psychological measurement.[5] Test users often speak of the "psychometric soundness" of tests, two key aspects of which are reliability and validity.

Reliability A good test or, more generally, a good measuring tool or instrument is *reliable.* As we will explain in Chapter 5, the criterion of reliability has to do with the *consistency* of the measuring tool, the precision with which the test measures, and the extent to which error is present in measurements. In theory, the perfectly reliable measuring tool consistently measures in the same way. For example, to determine if a digital scale was a reliable measuring tool, we might take repeated measures of the same standard weight, such as a 1-pound gold bar. If the scale repeatedly indicated that the gold bar weighed 1 pound, we would say that the scale was a reliable measuring instrument. If another scale repeatedly indicated that the gold bar weighed exactly 1.3 pounds, we would still say that the scale was reliable (although inaccurate and invalid), because the scale provided a consistent result. But suppose we weighed the bar ten times and six of those times the scale registered 1 pound, on two occasions the bar weighed in at a fraction of an ounce less than a pound, and on two other occasions it weighed in at a fraction of an ounce more than a pound . . . would the scale still be considered a reliable instrument?

Whether we are measuring gold bars, behavior, or anything else, unreliable measurement is a problem to avoid. We want to be reasonably certain that the measuring tool or test we are using will yield the same numerical measurement every time we observe the same thing under the same conditions. Psychological tests, like other tests and instruments, are reliable to varying degrees. Specific procedures for making determinations as to the reliability of an instrument will be introduced in Chapter 5, as will the various types of reliability.

Validity A good test is a *valid* test, and a test is considered to be valid for a particular purpose if it in fact measures what it purports to measure. In the gold bar example cited earlier, the scale that consistently indicated that the 1-pound gold bar did, in fact, weigh 1 pound is a valid scale. Likewise, a test of reaction time is a valid test if it truly measures reaction time. A test of intelligence is a valid test if it truly measures intelligence. A potential problem, however, is that although there is relatively little controversy about the definition of a term such as reaction time, a great deal of controversy exists about the definition of intelligence. The validity of a particular test might be questioned with regard to the definition of whatever that test purports to measure. A test creator's conception of what constitutes intelligence might be different from someone else's, and therein lies the basis for a claim that the test is invalid.

Questions regarding a test's validity may focus on the items that collectively make up the test. Do the items adequately sample the range of areas that must be sampled to adequately measure the construct? Individual items will also come under scrutiny in an investigation of a test's validity; how do individual items contribute to or take away from the test's validity? The validity of a test may also be questioned in regard to the scores derived from an administration of the test; what do the scores really tell us about the targeted construct? How are high and low scores on the test related to testtakers' behavior? In general, how do scores on this test relate to scores on other tests purporting

5. Variants of these words include the adjective *psychometric* and the nouns *psychometrist* and *psychometrician.* Traditionally, a **psychometrist** holds a master's degree and is qualified to administer specific tests. A **psychometrician** holds a doctoral degree in psychology or some related field (such as education) and specializes in areas such as individual differences, quantitative psychology, or theory of assessment.

to measure the same construct? How do scores on this test relate to scores on other tests purporting to measure opposite types of constructs? For example, we might expect one person's score on a valid test of introversion to be inversely related to that same person's score on a valid test of extraversion. That is, the higher the introversion test score, the lower the extraversion test score, and vice versa.

As we will see when we discuss validity in greater detail in Chapter 6, questions concerning the validity of a particular test or assessment procedure extend beyond the specific test or procedure per se. Critical validity-related questions concern the way in which data from a particular test or assessment procedure are used.

Other considerations If the purpose of a test is to compare the performance of the testtaker with the performance of other testtakers, a good test is one that contains adequate **norms.** Also referred to as *normative data,* norms provide a standard with which the results of measurement can be compared. These types of tests are referred to as **norm-referenced,** and a common goal of such tests is to yield information on the testtaker's standing or ranking relative to some comparison group of testtakers. The SAT and the GRE are two examples of norm-referenced tests; scores reflect the testtaker's standing relative to other testtakers. As an aid to a prospective test user in judging the appropriateness of administering, scoring, and interpreting a norm-referenced test, a complete description of the **norm group** or **normative sample** (the people who were tested with the instrument and with whom current testtakers' performance is being compared) is required. Unfortunately, manuals for norm-referenced tests differ widely in the specificity they employ in describing the norm group. Because of its greater specificity, a description such as "200 male, Hispanic, freshman community college students between the ages of 18 and 20 at New York City Community College" is preferable to one such as "many minority college students from a large community college in the East." In general, the closer the match between the norm group and the examinee(s), the more appropriate the test may be for a given purpose. Some norm-referenced tests are better than others because of the size of the normative sample; all other things being equal, the larger the normative sample, the better.

In contrast to norm-referenced tests, some tests, particularly in the fields of educational and industrial or organizational assessment, are **criterion-referenced.**[6] Whereas norm-referenced tests yield information about a testtaker's relative standing, criterion-referenced tests yield information about an individual's mastery of a particular skill. Has this applicant mastered the skills necessary to be a pilot for this airline? Has this student mastered the ability to spell "sand"? Has this group home member mastered the skills necessary for independent living? These are the types of questions criterion-referenced tests may seek to answer. When evaluating a criterion-referenced test, key issues concern the definition of the criterion used by the test developer, the relevance of the test's criterion to the objectives of the current assessment, and the evidence in hand that supports the use of the test for the contemplated purpose (see *Everyday Psychometrics.*)

6. As we will point out in Chapter 4, the criterion-referenced approach has been referred to of late with various terminology such as "domain-referenced," "content-referenced" and "objective-referenced." Our view is that this plethora of alternate terminology, although offered in the spirit of precision, tends to muddle rather than clarify distinctions between norm-referenced and criterion-referenced approaches. Assessment with reference to a criterion has traditionally been associated with the assessment of learning outcomes as opposed to mere content (or domain). Terms such as "content-referenced" and "domain-referenced" speak more to learning content (or domain) than to a learning outcome. Further, and what can be so confusing, is that a norm-referenced test may be content- or domain-referenced in the sense that it is linked or referenced to a particular content area; still, only the term "criterion-referenced" is used as if synonymous with "content-" or "domain-referenced."

EVERYDAY PSYCHOMETRICS

Putting Tests to the Test

For experts in the field of testing and assessment, a number of questions occur almost reflexively when evaluating a test or measurement technique. You may not be an assessment expert yet, but your consideration of questions such as the following will represent a significant first step in that direction. Try to think of these questions when you come across mention of various tests in this book, in other books and journal articles, and in life. These questions will help you evaluate the psychometric soundness of tests and other measurement methods.

Why Use This Particular Instrument or Method?

A choice of measuring instruments typically exists when it comes to measuring a particular psychological or educational variable, and the test user must therefore choose from many available tools. Published information, such as test catalogues, test manuals, and published test reviews, can be of great value in coming to a decision regarding the use of a particular test. Unpublished sources of information, such as information obtained by writing directly to the test developer or test publisher, may also be a possibility. Some of the questions the prospective test user will raise relate to the objectives of the test and the goodness of fit between those objectives and the objectives of the testing or assessment. What type of information will result from an administration of this test? Do alternate forms of this test exist and, if so, how might they be used? How long does it take to administer this test? What is the recommended age range for test-takers and what reading level is required? How will this resulting information be applied to answer the test referral question? What types of decisions can or cannot be made on the basis of information from the use of this test? What other information will be required in order to adequately answer the test referral question?

Are There Any Published Guidelines Relevant to the Use of This Test?

Measurement professionals make it their business to be aware of published guidelines from professional associations and related organizations relevant to the use of tests and measurement techniques. So, for example, suppose you are a psychologist called upon to provide input to a court in the matter of a child custody decision. More specifically, the court has asked you for your professional opinion regarding the parenting capacity of one parent. How would you proceed? Many psychologists who perform such evaluations use a psychological test as part of the evaluation process. However, the psychologist performing such an evaluation is—or should be—aware of the guidelines promulgated by the American Psychological Association's Committee on Professional Practice and Standards (1994). These guidelines describe three types of assessments relevant to a child custody decision: (1) the assessment of parenting capacity, (2) the assessment of psychological and developmental needs of the child, and (3) the assessment of the goodness of fit between the parent's capacity and the child's needs. Clearly, an evaluation of a parent, or even two parents, does not provide the evaluator with sufficient information to express an opinion as to custody. Rather, only an evaluation of the parents (or others seeking custody), the child, and the goodness of fit between the needs and capacity of each of the parties can provide information relevant to an educated opinion about child custody.

There are many psychological tests and measurement procedures used to obtain information about parenting capacity (Holden & Edwards, 1989; Lovejoy et al., 1999; Touliatos et al., 1991). According to Heinze and Grisso (1996), some of the most commonly used instruments are the Ackerman-Schoendorf Scales for Parent Evaluation of Custody, the Bricklin Perceptual Scales, the Bricklin Perception of Relationships Test, the Child Abuse Potential Inventory (CAP), the Parent-Child Relationship Inventory, and the Parenting Stress Index (PSI). Regardless of the particular test(s) employed, the psychologist will use other sources of data, such as interviews, behavioral observation, and document analysis, in the evaluation of parenting capacity. This is consistent both with accepted professional practice as well as the published guideline that encourages psychologists to "use multiple methods of data gathering" (APA, 1994a, p. 679). Data from multiple sources of data can have the effect of providing varied sources of support for a professional opinion, conclusion, or recommendation.

The area of child custody evaluation provides a useful illustration of why mere knowledge of assessment or of a test may not adequately equip an assessor to assess. Assessors who undertake child custody evaluations must have working familiarity not only with the specific tools they use and the current literature in psychological assessment in general, but with the ever-changing laws and professional guidelines applicable to such evaluations, as well as the current literature in areas such as child development, family dynamics, and divorce. Executing a competent child custody

(continued)

EVERYDAY PSYCHOMETRICS

Putting Tests to the Test *(continued)*

evaluation is no simple matter, and there are many published resources designed to assist professionals who wish to become involved in this type of work (for example, Ackerman, 1995; Bushard & Howard, 1994; Schultz et al., 1989; Stahl, 1995).

Is This Instrument Reliable?

Earlier, we introduced you to the psychometric concept of reliability and noted that it had to do with the consistency of measurement. Here, we hope to pique your interest in learning more about this concept by pointing out that measuring reliability is not always a straightforward matter. As an example, consider one of the tests that might be used in the evaluation of parenting capacity, the Bricklin Perceptual Scales (BPS; Bricklin, 1984). The BPS was designed to explore a child's perception of father and mother. A measure of one type of reliability, referred to as test-retest reliability, would indicate how consistent a child's perception of father and mother is over time. However, the BPS test manual contains no reliability data because as Bricklin (1984, p. 42) put it, "there are no reasons to expect the measurements reported here to exhibit any particular degree of stability, since they should vary in accordance with changes in the child's perceptions." Such an assertion has not stopped others (such as Speth, 1992) from exploring the test-retest reliability of the BPS. But whether or not one accepts Bricklin's assertion regarding the need for reliability data, such opinions illustrate the complexity of reliability questions—as well as the need for multiple sources of data to strengthen arguments regarding the confirmation or rejection of a hypothesis.

Is This Instrument Valid?

Validity, as you have learned, refers to the extent that a test measures what it purports to measure. Like reliability, questions related to the validity of a test can be complex and colored more in shades of gray than black or white. So, for example, even if data from a test such as the BPS were valid for the purpose of gauging children's perceptions of their parents, the data would not necessarily be valid as the sole source on which to base an opinion regarding child custody (Brodzinsky, 1993). In this context, Heinze and Grisso (1996) bemoaned what they saw as a trend by experts to rely on data concerning perceptions of the desirability of parents:

> Questions of parental desirability cannot be answered without reference to the characteristics, needs, and demands of the specific child who is in need of parenting. We suspect that no instrument that only assesses parents (e.g., whether through children's perceptions or direct observations of parents themselves) can ever meet basic scientific standards for making judgments about "preferred parents," or for making comparisons between parents that would justify suggesting that one parent's abilities are more desirable than the other's. (p. 310)

Instruments designed to measure variables such as stressful reactions to parenting (such as the PSI) and the potential for child abuse (such as the CAP) have yielded valuable data that could be very useful to courts as they evaluate all of the elements necessary for an informed judgment in child custody matters (Heinze & Grisso, 1996). However, in the courtroom and beyond, questions concern-

Must you be an assessment expert in order to be able to know a good test when you see one? Not necessarily. In some cases, all you need is to be good at retrieving relevant information about a particular test. In many instances, such information is as close as your university library and as available as cyberspace.

Reference Sources for Test Information

Many reference sources exist for learning more about published tests. These sources vary with respect to detail; some merely provide descriptions of tests, whereas others provide very technical information regarding reliability, validity, norms, and other such matters.

ing which test or combination of tests is valid for what purpose under what conditions sometimes stimulate heated debate and controversy.

What Inferences May Reasonably Be Made from This Test Score and How Generalizable Are the Findings?

The *raison d'être* (or *reason for being*) of many psychological tests and other tools of psychological assessment is to make inferences about behavior. In evaluating a test, it is therefore critical to consider the inferences that may reasonably be made as a result of administering that test. Will we learn something about a child's readiness to begin first grade? How prepared a student is for the first year of college at a particular institution? Whether the odds favor success for an independent life outside an institution for a person with a disability? Whether one is harmful to oneself or others to the extent that involuntary institutionalization is required? These represent but a small sampling of critical questions for which answers must be inferred on the basis of test scores and other data derived from various tools of assessment.

Intimately related to considerations regarding the inferences that can be made are considerations regarding the generalizability of the findings. Even from our brief introduction of the subject of norms, you are probably aware that normative data provide a context in which to interpret and generalize from test results. And following the discussion above regarding the complexity of measuring reliability and validity, you may have (correctly) anticipated comments about the complexity of gauging the generalizability of test findings. Consider, for example, that the normative sample for the Parenting Stress Index (PSI) consisted of 2,633 parents, drawn primarily from the state of Virginia. The majority of the children in the sample were under 5 years of age and Caucasian. How generalizable would you say the findings from an administration of the PSI are to non-Caucasian parents? If this is a question that occurred to you, you are in good company (see, for example, Krauss, 1993; McBride, 1989; Teplin et al., 1991; Younger, 1991). In fact, adaptations of the PSI have been made to include parents from different cultures (Abidin, 1990; Beebe et al., 1993; Black et al., 1993).

In addition to issues regarding the applicability of the norms, a number of other factors may give rise to questions regarding the generalizability of a test or a particular administration of a test. The wording of test items, for example, may have the effect of biasing scores in some way. So, for example, it may be that all other things being equal, the BPS is biased toward more favorable perceptions of mothers. Mothers and fathers may score similarly on all of the subtests except the Supportiveness subscale on which mothers tend to score higher (Heinze & Grisso, 1996).

The question of generalizability of findings may also be raised with regard to issues concerning a particular administration of a test. Most published tests have very explicit directions that test administrators—or a computer, if the test is computer-administered—must follow to the letter. If test administration is compromised in any way—whether by design, negligence, or any other reason—the generalizability of the data derived from the testing has also been compromised.

And so, although you may not yet be an expert in measurement, you are now armed with a working knowledge of the types of questions such experts ask when evaluating any test or measurement technique.

Test manuals Detailed information concerning the development of a particular test, the normative sample, the test's reliability and validity, and other such information should be found in the manual for the test itself. The chances are good that somewhere within your university (be it the library or the counseling center), a collection of popular psychological test manuals is maintained. If not, most test publishers are willing to sell a test manual by itself, sometimes within some sort of sampler kit.

Test catalogues Perhaps one of the most readily accessible sources of information about a test is a catalogue distributed by the publisher of the test. Because most test publishers make available catalogues of their offerings, this source of test information can be tapped

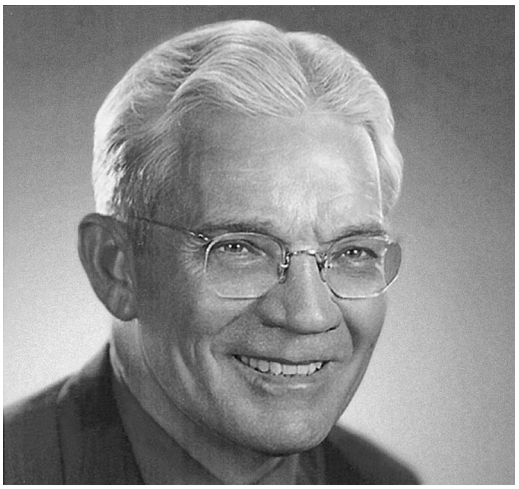

Figure 1–6
Oscar Krisen Buros (1906–1978)

Buros is best remembered for being the creator of the Mental Measurements Yearbook *(MMY), a kind of* Consumer Reports *for tests and a much needed source of "psychometric policing" (Peterson, 1997, p. 718). His work lives on at the Buros Institute of Mental Measurements at the University of Nebraska, Lincoln. In addition to the MMY, which is updated periodically, the institute publishes a variety of other test-related publications.*

by a simple telephone call, e-mail, or note. As you might expect, however, publishers' catalogues usually contain only a brief description of the test and seldom contain the kind of detailed technical information that a prospective user of the test might require. Further, the objective of the catalogue is to sell the test. Expect any quotations from reviews critical of the test to be excluded from the description.

Reference volumes The Buros Institute of Mental Measurements provides "one-stop shopping" for a great deal of test-related information including lists of test publishers and recently published or newly revised tests, as well as test reviews. The initial version of what would evolve into the *Mental Measurements Yearbook* (MMY) was compiled by Oscar Buros (Figure 1–6) as early as 1933. At this writing, the latest edition of this authoritative compilation of test reviews is the *13th Mental Measurements Yearbook* (Impara & Plake, 1998), although the *14th* cannot be far behind. The Buros Institute also publishes *Tests in Print* as well as a number of other test-related reference works. For a list of its latest offerings, as well as links to a number of other useful test-related test databases, visit the Institute's Web site at *http://www.unl.edu/buros/.*

Journal articles Articles relevant to the development and use of sundry tests and measurement methods can be found in the pages of a wide array of behavioral science journals (such as *Psychological Bulletin, Psychological Review, Professional Psychology: Research and Practice,* and *Journal of Personality and Social Psychology*), as well as journals that focus more specifically on matters related to testing and assessment (such as *Psychological Assessment, Educational and Psychological Measurement, Applied Measurement in Education,* and the *Journal of Personality Assessment*). Journals such as *Psychology, Public Policy, and Law* and *Law and Human Behavior* frequently contain highly informative articles on legal and ethical issues and controversies as they relate to psychological testing and assessment.

In addition to articles relevant to specific tests, journals are a rich source of information regarding important trends in testing and assessment. For example, with reference to clinical psychological assessment, the negative impact of managed health care and the reluctance or refusal of insurers to pay for assessment services has spurred a great deal of self-evaluation on the part of those in the business of evaluation (Acklin, 1996; Backlar, 1996; Camara et al., 2000; Eisman et al., 1998; Miller, 1996; Piotrowski et al., 1998). While

Table 1–1
Some Internet Web Site Addresses for Test Publishers

Academic Therapy www.academictherapy.com	James Stanfield Company www.stanfield.com	Pro-Ed www.proedinc.com
American Guidance Service www.agsnet.com	Lafayette Instruments www.licmef.com	Riverside Publishing www.riverpub.com
Consulting Psychologists Press www.cpp-db.com	Meritech, Inc. www.meritech.com	Scholastic Testing Service www.ststesting.com
CTB McGraw-Hill www.ctb.com	Multi-Health Systems www.mhs.com	Slosson Educational Publications www.slosson.com
Educator's Publishing Service www.epsbooks.com	National Computer Assessments www.ncs.com	Sopris West www.sopriswest.com
Harcourt Brace Educational Measurement www.hbem.com	Psychological Assessment Resources www.parinc.com	Stoelting www.stoeltingco.com
Institute for Personality and Ability Testing www.ipat.com	The Psychological Corporation www.psychcorp.com	Vort www.vort.com

critics of clinical assessment argue that testing and assessment is too expensive, too time consuming, and of too little value (Griffith, 1997), more informed reviews of the issues find abundant empirical support for the value of the enterprise (Kubiszyn et al., 2000).

Online databases The American Psychological Association (APA) maintains a number of databases useful in locating psychology-related information in journal articles, book chapters, and doctoral dissertations. PsycINFO is a database of abstracts dating back to 1887. ClinPSYC is a database derived from **PsycINFO** that focuses on abstracts of a clinical nature. PsycSCAN: Psychopharmacology contains abstracts of articles having to do with psychopharmacology. PsycARTICLES is a database of full-length articles dating back to 1988. **PsycLAW** is a free database available to everyone that contains discussions of selected topics having to do with psychology and law. It can be accessed at *http://www.psychlaw.org.* For more information on any of these databases, visit APA's Web site at *http://www.apa.org.*

Educational Testing Service (ETS), "the world's largest and most influential testing organization" (Frantz & Nordheimer, 1997), maintains its own Web site at *http://www.ets.org.* The site contains a wealth of information about college and graduate school admission and placement tests, as well as many related resources. If you wanted to try your hand at some practice questions for a test such as the Graduate Record Examination (GRE), for example, this is the place to go. For more information, ETS can be contacted by e-mail at *etsinfo@ets.org.* A list of Web sites for publishers of other educational and psychological tests is presented in Table 1–1.

Other sources Your school library contains a number of other sources that may be used to acquire information about tests and test-related topics. For example, two sources for exploring the world of unpublished tests and measures are the *Directory of Unpublished Experimental Measures* (Goldman & Mitchell, 1995) and *Tests in Microfiche* available from Test Collections. APA makes available *Finding Information About Psychological Tests* (1995), its own guide to locating test-related information.

Armed with a wealth of background information about tests and other tools of assessment, let's explore various historical, cultural, and legal/ethical aspects of the assessment enterprise.

Self-Assessment

Test your understanding of elements of this chapter by seeing if you can explain each of the following terms, expressions, and abbreviations:

ABAP
ABPP
alternate assessment
assessment
behavioral observation
case history data
construct
diagnosis
diagnostic test
erg
ergonomics
error
error variance
format
interview
measurement
MMY
norms
portfolio
protocol
PsycINFO
psychological assessment
psychological autopsy
psychological test
psychological testing
psychometrics
psychometry
Public Law 94-142
Public Law 99-457
rapport
reliability
role play test
scale
scaling
score
scoring
standard error of measurement
state
test
test catalogue
test developer
testing
test manual
testtaker
test user
trait
validity

The same words or terms may appear at the end of different chapters. So, for example, you will see words such as "norms," "reliability," and "validity" at the end of some succeeding chapters—at which point you may be able to provide a more detailed explanation of what these words mean.

Another aid to self-assessment and learning is the crossword puzzles presented in the companion study guide and workbook to this textbook, *Exercises in Psychological Testing and Assessment* (Cohen, 2002). Each chapter begins with a puzzle that contains "clues" to key concepts, terms, and/ or names presented in the chapter.