One

INTRODUCTION TO PSYCHOLOGICAL TESTS AND THEIR USES

The first and most general meaning of the term *test* listed in the dictionary is "a critical examination, observation, or evaluation." Its closest synonym is *trial*. The word *critical,* in turn, is defined as "relating to . . . a turning point or specially important juncture" (*Merriam-Webster's Collegiate Dictionary,* 1995). No wonder, then, that when the term *psychological* appears in front of the word *test,* the resulting phrase acquires a somewhat threatening connotation. Psychological tests are often used to evaluate individuals at some turning point or significant juncture in their lives. Yet, in the eyes of many people, tests seem to be trials on which too much depends and about which they know all too little. To a large extent, the purpose of this book is to give readers enough information about psychological tests and testing to remove their threatening connotations and to provide the means whereby consumers of psychological tests can gain more knowledge about their specific uses.

Thousands of instruments can accurately be called *psychological tests.* Many more usurp the label either explicitly or by suggestion. The first objective of this book is to explain how to separate the former from the latter. Therefore, we start with the defining features that legitimate psychological tests of all types share. These features not only define psychological tests but also differentiate them from other kinds of instruments.

PSYCHOLOGICAL TESTS

A *psychological test* is a systematic procedure for obtaining samples of behavior, relevant to cognitive or affective functioning, and for scoring and evaluating those samples according to standards. A clarification of each of the main terms in this definition is vital to an understanding of all future discussion of tests. Rapid Reference 1.1 explains the meaning and rationale of all the elements in the definition of a psychological test. Unless every condition mentioned in the definition is met, the procedure in question cannot accurately be called a psychological test. It is,

Rapid Reference 1.1

Basic Elements of the Definition of Psychological Tests

Defining Element	Explanation	Rationale
Psychological tests are *systematic* procedures.	They are characterized by planning, uniformity, and thoroughness.	Tests *must be* demonstrably objective and fair to be of use.
Psychological tests are *samples of behavior.*	They are small subsets of a much larger whole.	Sampling behavior is efficient because the time available is usually limited.
The behaviors sampled by tests are *relevant to cognitive or affective functioning* or both.	The samples are selected for their empirical or practical psychological significance.	Tests, unlike mental games, exist to be of use; they are tools.
Test results are *evaluated* and *scored.*	Some numerical or category system is applied to test results, according to preestablished rules.	There should be no question about what the results of tests are.
To evaluate test results it is necessary to have *standards* based on empirical data.	There has to be a way of applying a common yardstick or criterion to test results.	The standards used to evaluate test results lend the only meaning those results have.

however, important to remember that in essence, psychological tests are simply behavior samples. Everything else is based on inferences.

Psychological tests are often described as *standardized* for two reasons, both of which address the need for objectivity in the testing process. The first has to do with uniformity of procedure in all important aspects of the administration, scoring, and interpretation of tests. Naturally, the time and place when a test is administered, as well as the circumstances under which it is administered and the examiner who administers it, affect test results. However, the purpose of standardizing test procedures is to make all the variables that are under the control of the examiner as uniform as possible, so that everyone who takes the test will be taking it in the same way.

The second meaning of standardization concerns the use of standards for evaluating test results. These standards are most often norms derived from a group of individuals—known as the *normative* or *standardization sample*—in the process of developing the test. The collective performance of the standardization group or groups, both in terms of averages and variability, is tabulated and be-

comes the standard against which the performance of other individuals who take the test after it is standardized will be gauged.

Strictly speaking, the term *test* should be used only for those procedures in which test takers' responses are evaluated based on their correctness or quality. Such instruments always involve the appraisal of some aspect of a person's cognitive functioning, knowledge, skills, or abilities. On the other hand, instruments whose responses are neither evaluated nor scored as right-wrong or pass-fail are called *inventories, questionnaires, surveys, checklists, schedules,* or *projective techniques,* and are usually grouped under the rubric of *personality tests.* These are tools designed to elicit information about a person's motivations, preferences, attitudes, interests, opinions, emotional makeup, and characteristic reactions to people, situations, and other stimuli. Typically, they use questions of the multiple-choice or true-false type, except for projective techniques, which are open ended. They can also involve making forced choices between statements representing contrasting alternatives, or rating the degree to which one agrees or disagrees with various statements. Most of the time personality inventories, questionnaires, and other such instruments are of the self-report variety but some are also designed to elicit reports from individuals other than the person being evaluated (e.g., a parent, spouse, or teacher). For the sake of expediency, and following common usage, the term *test* will be used throughout this book to refer to all instruments, regardless of type, that fit the definition of a psychological test. Tests that sample knowledge, skills, or cognitive functions will be designated as *ability tests,* whereas all others will be referred to as *personality tests.*

DON'T FORGET

- The word *test* has multiple meanings.
- The term *psychological test* has a very specific meaning.
- In this book, *test* will be used to refer to all instruments that fit the definition of *psychological test.*
- Tests designed to sample skills, knowledge, or any other cognitive function will be referred to as *ability tests;* all others will be labeled as *personality tests.*

Other Terms Used in Connection with Tests and Test Titles

Some other terms that are used, sometimes loosely, in connection with tests bear explaining. One of these is the word *scale,* which can refer to

- a whole test made up of several parts, for example, the *Stanford-Binet Intelligence Scale;*
- a subtest, or set of items within a test, that measures a distinct and spe-

cific characteristic, for example, the *Depression scale* of the Minnesota Multiphasic Personality Inventory (MMPI);
- an array of subtests that share some common characteristic, for example, the *Verbal scales* of the Wechsler intelligence tests;
- a separate instrument made up of items designed to evaluate a single characteristic, for example, the *Internal-External Locus of Control Scale* (Rotter, 1966); or
- the numerical system used to rate or to report value on some measured dimension, for example, a *scale* ranging from 1 to 5, with 1 meaning *strongly disagree* and 5 *strongly agree*.

Thus, when used in reference to psychological tests, the term *scale* has become ambiguous and lacking in precision. However, in the field of psychological measurement—also known as *psychometrics*—*scale* has a more precise meaning. It refers to a group of items that pertain to a single variable and are arranged in order of difficulty or intensity. The process of arriving at the sequencing of the items is called *scaling*.

Battery is another term often used in test titles. A battery is a group of several tests, or subtests, that are administered at one time to one person. When several tests are packaged together by a publisher to be used for a specific purpose, the word *battery* usually appears in the title and the entire group of tests is viewed as a single, whole instrument. Several examples of this usage occur in neuropsychological instruments (such as the Halstead-Reitan Neuropsychological Battery) where many cognitive functions need to be evaluated, by means of separate tests, in order to detect possible brain impairment. The term *battery* is also used to designate any group of individual tests specifically selected by a psychologist for use with a given client in an effort to answer a specific referral question, usually of a diagnostic nature.

PSYCHOLOGICAL TESTS AS TOOLS

The most basic fact about psychological tests is that they are tools. This means that they are always a means to an end and never an end in themselves. Like other tools, psychological tests can be exceedingly helpful—even irreplaceable—when used appropriately and skillfully. However, tests can also be misused in ways that may limit or thwart their usefulness and, at times, even result in harmful consequences.

A good way to illustrate the similarities between tests and other, simpler, tools

is the analogy between a test and a hammer. Both are tools for specific purposes, but can be used in a variety of ways. A hammer is designed basically for pounding nails into various surfaces. When used appropriately, skillfully, and for its intended purpose a hammer can help build a house, assemble a piece of furniture, hang pictures in a gallery, and do many other things. Psychological tests are tools designed to help in drawing inferences about individuals or groups. When tests are used appropriately and skillfully they can be key components in the practice and science of psychology.

DON'T FORGET

Psychological tests are evaluated at *two* distinct points and in *two* different ways:

1. When they are being considered as potential tools by prospective users; at this point, their technical qualities are of primary concern.
2. Once they are placed in use for a specific purpose; at this point, the skill of the user and the way tests are used are the primary considerations.

Just as hammers may be used for good purposes other than those for which they were intended (e.g., as paperweights or doorstops), psychological tests may also serve purposes other than those for which they were designed originally, such as increasing self-knowledge and self-understanding. Furthermore, just as hammers can hurt people and destroy things when used incompetently or maliciously, psychological tests can also be used in ways that do damage. When test results are misinterpreted or misused, they can harm people by labeling them in unjustified ways, unfairly denying them opportunities, or simply discouraging them.

All tools, be they hammers or tests, can be evaluated based on how well they are designed and built. When looked at from this point of view, prior to being used, tests are evaluated only in a limited, technical sense and their appraisal is of interest mostly to potential users. Once they are placed into use, however, tests cannot be evaluated apart from the skills of their users, the ways they are used, and the purposes for which they are used. This in-use evaluation often involves issues of policy, societal values, and even political priorities. It is in this context that the evaluation of the use of tests acquires practical significance for a wider range of audiences.

Testing Standards

Because of the unique importance of tests to all the professionals who use them and to the general public, since the mid-1950s, three major professional organi-

Rapid Reference 1.2

Testing Standards

- This designation will be used frequently throughout this book to refer to the *Standards for Educational and Psychological Testing,* published jointly in 1999 by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education.
- The *Testing Standards* are the single most important source of criteria for the evaluation of tests, testing practices, and the effects of test use.

zations have joined forces to promulgate standards that provide a basis for evaluating tests, testing practices, and the effects of test use. The most recent version of these is the *Standards for Educational and Psychological Testing*, published in 1999 by the American Educational Research Association (AERA) and prepared jointly by AERA, the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). As Rapid Reference 1.2 indicates, these standards are cited throughout this book and hereafter will be referred to as the *Testing Standards.*

PSYCHOLOGICAL TESTS AS PRODUCTS

The second most basic fact about psychological tests is that they are products. Although this is an obvious fact, most people are not mindful of it. Tests are products primarily marketed to and used by professional psychologists and educators, just as the tools of dentistry are marketed and sold to dentists. The public at large remains unaware of the commercial nature of psychological tests because they are advertised through publications and catalogs targeted to the professionals who use them. Nevertheless, the fact remains that many, if not most, psychological tests are conceived, developed, marketed, and sold for applied purposes in education, business, or mental health settings. They also must make a profit for those who produce them, just like any other commercial product.

As we will see, from the very beginning, the psychological testing enterprise was fueled principally by the need to make practical decisions about people. Since tests are professional tools that can be used both to benefit people *and* as commercial products, some clarification of the various parties in the testing enterprise and their roles is justified. Rapid Reference 1.3 shows a list of the major participants in the testing process and their roles.

As the *Testing Standards* stipulate, "the interests of the various parties involved in the testing process are usually, but not always, congruent" (AERA, APA, NCME, 1999, p. 1). For example, test *authors* are usually, though not always, aca-

Rapid Reference 1.3

Participants in the Testing Process and Their Roles

Participants	Their Roles in the Testing Process
Test authors and developers	They conceive, prepare, and develop tests. They also find a way to disseminate their tests, by publishing them either commercially or through professional publications such as books or periodicals.
Test publishers	They publish, market, and sell tests, thus controlling their distribution.
Test reviewers	They prepare evaluative critiques of tests based on their technical and practical merits.
Test users	They select or decide to take a specific test off the shelf and use it for some purpose. They may also participate in other roles, e.g., as examiners or scorers.
Test administrators or examiners	They administer the test either to one individual at a time or to groups.
Test takers	They take the test by choice or necessity.
Test scorers	They tally the raw responses of the test taker and transform them into test scores through objective or mechanical scoring or through the application of evaluative judgments.
Test score interpreters	They interpret test results to their ultimate consumers, who may be individual test takers or their relatives, other professionals, or organizations of various kinds.

demicians or investigators who are mainly interested in psychological theorizing or research, rather than in practical applications or profits. Test *users* are most interested in the appropriateness and utility of the tests they use for their own purposes, whereas test *publishers* are naturally inclined to consider the profit to be made from selling tests foremost. Furthermore, participants in the testing process may perform one or more of all the various roles described in Rapid Reference 1.3. Test users may administer, score, and interpret the results of tests they have selected or may delegate one or more of these functions to others under their supervision. Similarly, test publishers can, and often do, hire test developers to create instruments for which they think a market exists. Nevertheless, of all participants in the testing process, the *Testing Standards* assign "the ultimate responsibility for appropriate test use and interpretation" predominantly to the test user (p. 112).

HISTORY OF PSYCHOLOGICAL TESTING

Even though psychological tests can be used to explore and investigate a wide range of psychological variables, their most basic and typical use is as tools in making decisions about people. It is no coincidence that psychological tests as we know them today came into being in the early part of the 20th century. Prior to the rise of urban, industrial, democratic societies, there was little need for most people to make decisions about others, outside of those in their immediate families or close circle of acquaintances. In rural, agrarian, autocratic societies, major life decisions about individuals were largely made for them by parents, mentors, rulers and, above all, by the gender, class, place, and circumstances into which people were born. Nonetheless, well before the 20th century, there are several interesting precursors of modern psychological testing within a variety of cultures and contexts.

Antecedents of Modern Testing in the Occupational Realm

A perennial problem in any field of employment is the question of how to select the best possible people for a given job. The oldest known precursors of psychological testing are found precisely in this area, within the system of competitive examinations developed in the ancient Chinese empire to select meritorious individuals for government positions. This forerunner of modern personnel selection procedures dates back to approximately 200 B.C.E. and went through a number of transformations in its long history (Bowman, 1989). The Chinese civil service examinations encompassed demonstrations of proficiency in music, archery, and horsemanship, among other things, as well as written exams in subjects such as law, agriculture, and geography. Apparently, the impetus for the development of this enlightened system of human resource utilization—open to any individual who was recommended to the emperor by local authorities throughout the empire—was the fact that China did not have the sort of hereditary ruling classes that were common in Europe until the 20th century. The Chinese imperial examination system ended in 1905 and was replaced with selection based on university studies. In the meantime, however, that system served as an inspiration for the civil service exams developed in Britain in the 1850s, which, in turn, stimulated the creation of the U.S. Civil Service Examination in the 1860s (DuBois, 1970).

Antecedents of Modern Testing in the Field of Education

One of the most basic questions in any educational setting is how to ascertain that students have acquired the knowledge or expertise their teachers try to instill in

them. Thus, it is not surprising that the earliest use of testing within the realm of education occurred during the Middle Ages with the rise of the first universities in Europe in the 13th century. At about that time, university degrees came to be used as a means of certifying eligibility to teach, and formal oral examinations were devised to give candidates for degrees an opportunity to demonstrate their competence (DuBois, 1970). Little by little, the use of examinations spread to the secondary level of education and, as paper became cheaper and more available, written examinations replaced the oral exams in most educational settings. By the late 19th century, in both Europe and the United States, examinations were a well-established method of ascertaining who should be awarded university degrees as well as who would be able to exercise a profession, such as medicine or law.

Antecedents of Modern Testing in Clinical Psychology

Another fundamental human question that can be and has been addressed by means of psychological testing is the problem of differentiating the "normal" from the "abnormal" within the intellectual, emotional, and behavioral arenas. However, in contrast to the occupational or educational contexts where the bases on which decisions are made have traditionally been fairly clear, the realm of psychopathology remained shrouded in mystery and mysticism for a much longer period.

Several antecedents of psychological tests stem from the field of psychiatry (Bondy, 1974). Many of these early tests were developed in Germany in the second half of the 19th century, although some of them date from the early part of that century and stemmed from France. Almost invariably these instruments were devised for the express purpose of assessing the level of cognitive functioning of patients with various kinds of disorders such as mental retardation or brain damage. Among the behavior samples used in these early tests were questions concerning the meaning of proverbs and the differences or similarities between pairs of words, as well as memory tasks such as the repetition of digit series presented orally. Many of the techniques developed in the 19th century were ingenious and survived to be incorporated into modern tests that are still in wide use (see McReynolds, 1986).

In spite of their cleverness, developers of the early forerunners of clinical tests were handicapped by at least two factors. One was the dearth of knowledge—and the abundance of superstitions and misconceptions—concerning psychopathology. In this regard, for instance, the distinction between psychosis and mental retardation was not even clearly formulated until 1838, when the French psychiatrist Esquirol suggested that the ability to use language is the most dependable

criterion for establishing a person's level of mental functioning. A second factor preventing the widespread dissemination and use of the early psychiatric tests was their lack of standardization in terms of procedures or of a uniform frame of reference against which to interpret results. To a large extent, the techniques developed by 19th-century neurologists and psychiatrists like Guislain, Snell, von Grashey, Rieger, and others were devised for the purpose of examining a specific patient or patient population. These behavior samples were collected in an unsystematic fashion and were interpreted by clinicians on the basis of their professional judgment rather than with reference to normative data (Bondy, 1974).

A significant breakthrough was achieved in psychiatry during the 1890s, when Emil Kraepelin set out to classify mental disorders according to their causes, symptoms, and courses. Kraepelin wanted to bring the scientific method to bear on psychiatry and was instrumental in delineating the clinical picture of schizophrenia and bipolar disorder, which—at the time—were known respectively as *dementia praecox* and *manic-depressive psychosis*. He proposed a system for comparing sane and insane individuals on the basis of characteristics such as distractibility, sensitivity, and memory capacity and even pioneered the use of the free-association technique with psychiatric patients. Although some of Kraepelin's students devised a battery of tests and continued to pursue the goals he had set out, the results of their work were not as fruitful as they had hoped (DuBois, 1970).

Antecedents of Modern Testing in Scientific Psychology

The investigations of the German psychophysicists Weber and Fechner in the mid-19th century initiated a series of developments that culminated in Wilhelm Wundt's creation of the first laboratory dedicated to research of a purely psychological nature in Leipzig, Germany, in 1879. This event is considered by many as the beginning of psychology as a separate, formal discipline, apart from philosophy. With the rise of the new discipline of experimental psychology, there also arose much interest in developing apparatus and standardized procedures for mapping out the range of human capabilities in the realm of sensation and perception. The first experimental psychologists were interested in discovering general laws governing the relationship between the physical and psychological worlds. They had little or no interest in individual differences—the main item of interest in differential psychology and psychological testing—which they, in fact, tended to view as a source of error. Nevertheless, their emphases on the need for accuracy in their measurements and for standardized conditions in the lab would prove to be important contributions to the forthcoming field of psychological testing.

Wundt's German lab flourished in the last decades of the 19th century and trained many psychologists from the United States and elsewhere who would go back to their countries to establish their own similar labs. At about the same time, an Englishman named Francis Galton became interested in the measurement of psychological functions from an entirely different perspective. Galton was a man of great intellectual curiosity and many accomplishments, whose privileged social and financial position allowed him to pursue a wide range of interests. He was also a cousin and a great admirer of Charles Darwin, whose theory of evolution of species by natural selection had revolutionized the life sciences in the mid-19th century. After reading his cousin's treatise on the origin of species, Galton decided to pursue his interest in the notion that intellectual gifts tend to run in families. To this end he set up an anthropometric lab in London, where for several years he collected data on a number of physical and physiological characteristics—such as arm span, height, weight, vital capacity, strength of grip, and sensory acuity of various kinds—on thousands of individuals and families. Galton was convinced that intellectual ability was a function of the keenness of one's senses in perceiving and discriminating stimuli, which he in turn believed was hereditary in nature. Through the accumulation and cross-tabulation of his anthropometric data, Galton hoped to establish both the range of variation in these characteristics, as well as their interrelationships and concordance across individuals with different degrees of familial ties (Fancher, 1996).

Galton did not succeed in his ultimate objective, which was to promote *eugenics,* a field of endeavor he had originated that aimed at improving the human race through selective breeding of its ablest specimens. To this end, he wanted to devise a way of assessing the intellectual capacity of children and adolescents through tests so as to identify the most gifted individuals early and encourage them to produce many offspring. Nevertheless, Galton's work was continued and considerably extended in the United States by James McKeen Cattell, who also tried, fruitlessly, to link various measures of simple discriminative, perceptive, and associative power (which he labeled "mental" tests) to independent estimates of intellectual level, such as school grades.

In light of some events of the 20th century, such as those in Nazi Germany, Galton's aim seems morally offensive to most contemporary sensibilities. However, at the time he coined the term *eugenics* and enunciated its aims, the genocidal potential of this endeavor was not generally perceived, and many illustrious individuals of that era were enthusiastic eugenicists. In the process of his pursuit, however misguided it may seem to us today, Galton did make significant contributions to the fields of statistics and psychological measurements. While charting data comparing parents and their offspring, for instance, he discovered the

phenomena of regression and correlation, which provided the groundwork for much subsequent psychological research and data analyses. He also invented devices for the measurement of hearing acuity and weight discrimination, and initiated the use of questionnaires and word association in psychological research. As if these accomplishments were not enough, Galton also pioneered the twin-study method that, once refined, would become a primary research tool in behavior genetics.

One additional contribution to the nascent field of psychological testing in the late 1800s deserves mention because it would lead directly to the first successful instrument of the modern era of testing. While studying the effects of fatigue on children's mental ability, the German psychologist Hermann Ebbinghaus—best known for his groundbreaking research in the field of memory—devised a technique known as the Ebbinghaus Completion Test. This technique called for children to fill in the blanks in text passages from which words or word-fragments had been omitted. The significance of this method, which would later be adapted for a variety of different purposes, is twofold. First, because it was given to whole classes of children simultaneously, it foreshadowed the development of group tests. What is more important, however, is that the technique proved to be an effective gauge of intellectual ability, as the scores derived from it corresponded well with the students' mental ability as determined by rank in class. As a result of this, Alfred Binet was inspired to use the completion technique and other complex mental tasks in developing the scale that would become the first successful intelligence test (DuBois, 1970).

The Rise of Modern Psychological Testing

By the early 1900s everything necessary for the rise of the first truly modern and successful psychological tests was in place:

- Laboratory tests and tools generated by the early experimental psychologists in Germany,
- Measurement instruments and statistical techniques developed by Galton and his students for the collection and analysis of data on individual differences, and
- An accretion of significant findings in the budding sciences of psychology, psychiatry, and neurology.

All of these developments provided the foundation for the rise of modern testing. The actual impetus for it, however, came from the practical need to make decisions in educational placement.

In 1904, the French psychologist Alfred Binet was appointed to a commission charged with devising a method for evaluating children who, due to mental retardation or other developmental delays, could not profit from regular classes in the public school system and would require special education. Binet was particularly well prepared for this task, as he had been engaged in investigating individual differences by means of a variety of physical and physiological measures, as well as tests of more complex mental processes, such as memory and verbal comprehension. In 1905, Binet and his collaborator, Theodore Simon, published the first useful instrument for the measurement of general cognitive abilities or global intelligence. The 1905 Binet-Simon scale, as it came to be known, was a series of 30 tests or tasks varied in content and difficulty, designed mostly to assess judgment and reasoning ability irrespective of school learning. It included questions dealing with vocabulary, comprehension, differences between pairs of concepts, and so on, as well as tasks that included repeating series of numbers, following directions, completing fragmentary text passages, and drawing.

The Binet-Simon scale was successful because it combined features of earlier instruments in a novel and systematic fashion. It was more comprehensive in its coverage than earlier instruments devoted to evaluating narrower abilities. It was, in fact, a small battery of carefully selected tests arranged in order of difficulty and accompanied by precise instructions on how to administer and interpret it. Binet and Simon administered the scale to 50 normal children ranging in age from 3 to 11 years, as well as to children with various degrees of mental retardation. The results of these studies proved that they had devised a procedure for sampling cognitive functioning whereby a child's general level of intellectual ability could be described quantitatively, in terms of the age level to which her or his performance on the scale corresponded. The need for such a tool was so acute that the 1905 scale would be quickly translated into other languages and adapted for use outside France.

The Birth of the IQ

Binet himself revised, expanded, and refined his first scale in 1908 and 1911. Its scoring developed into a system in which credit for items passed was given in terms of years and months so that a *mental level* could be calculated to represent quality of performance. In 1911 a German psychologist named William Stern proposed that the mental level attained on the Binet-Simon scale, relabeled as a *mental age score,* be divided by the chronological age of the subject to obtain a mental quotient that would more accurately represent ability at different ages. To eliminate the decimal, the mental quotient was multiplied by 100, and soon became known as the *intelligence quotient,* or *IQ*. This now-familiar score, a *true ratio*

IQ, was popularized through its use in the most famous revision of the Binet-Simon scales—the Stanford-Binet Intelligence Scale—published in 1916 by Lewis Terman. In spite of several problems with the ratio IQ, its use would last for several decades, until a better way of integrating age into the scoring of intelligence tests (described in Chapter 3) was devised by David Wechsler (Kaufman, 2000; Wechsler, 1939). Binet's basic idea—namely, that to be average, below average, or above average in intelligence means that one performs at, below, or above the level typical for one's age group on intelligence tests—has survived and become one of the primary ways in which intelligence is assessed.

While Binet was developing his scales in France, in England, Charles Spearman (a former student of Wundt's and follower of Galton) had been trying to prove empirically Galton's hypothesis concerning the link between intelligence and sensory acuity. In the process he had developed and expanded the use of correlational methods pioneered by Galton and Karl Pearson, and provided the conceptual foundation for *factor analysis,* a technique for reducing a large number of variables to a smaller set of factors that would become central to the advancement of testing and trait theory.

Spearman also devised a theory of intelligence that emphasized a general intelligence factor (or *g*) present in all intellectual activities (Spearman, 1904a, 1904b). He had been able to gather moderate support for Galton's notions by correlating teachers' ratings and grades with measures of sensory acuity, but soon realized that the tasks assembled in the Binet-Simon scale provided a far more useful and reliable way of assessing intelligence than the tools he had been using. Even though Spearman and Binet differed widely in their views about the nature of intelligence, their combined contributions are unsurpassed in propelling the development of psychological testing in the 20th century.

Group Testing

At the time Binet died, in 1911, he had already considered the possibility of adapting his scale to other uses and developing group tests that could be administered by one examiner to large groups for use in the military and other settings. The fulfillment of that idea, however, would not take place in France but in the United States, where the Binet-Simon scale had been rapidly translated and revised for use primarily with schoolchildren and for the same purpose as it had been developed in France.

Upon the entry of the United States into World War I in 1917, the APA president, Robert Yerkes, organized a committee of psychologists to help in the war effort. It was decided that the most practical contribution would be to develop a group test of intelligence that could be efficiently administered to all recruits

into the U.S. Army, to help in making personnel assignments. The committee, made up of leading test experts of the day, including Lewis Terman, hastily assembled and tried out a test that came to be known as the *Army Alpha*. It consisted of eight subtests measuring verbal, numerical, and reasoning abilities, as well as practical judgment and general information. The test, which would eventually be administered to more than a million recruits, made use of materials from various other instruments, including the Binet scales. In constructing it, the committee relied heavily on an unpublished prototype group test developed by Arthur Otis, who had devised multiple-choice items that could be scored objectively and rapidly.

The Army Alpha proved to be extremely useful. It was followed rapidly by the Army Beta, a supposedly equivalent test that did not require reading and could thus be used with recruits who were illiterate or non–English speaking. Unfortunately, the haste with which these tests were developed and put into use resulted in a number of inappropriate testing practices. In addition, unwarranted conclusions were made on the basis of the massive amounts of data that quickly accumulated (Fancher, 1985). Some of the negative consequences of the ways in which the Army testing program, and other massive testing efforts from that era, were implemented damaged the reputation of psychological testing in ways that have been difficult to surmount. Nevertheless, through the mistakes that were made early in the history of modern testing, a great deal was learned that later served to correct and improve the practices in this field. Furthermore, with the Army tests the field of psychology decisively stepped out of the lab and academic settings and demonstrated its enormous potential to contribute to real-world applications.

After World War I, psychological testing came into its own in the United States. Otis published his Group Intelligence Scale, the test that had served as a model for the Army Alpha, in 1918. E. L. Thorndike, another important American pioneer working at Teachers College at Columbia, produced an intelligence test for high school graduates, standardized on a more select sample (namely, college freshmen) in 1919. From then on, the number of published tests grew rapidly. Procedural refinements were also swiftly instituted in test administration and scoring. For example, test items of different types began to be presented in a mixed order rather than as separate subtests so that an overall time limit could be used for a test, eliminating the need for separate timing of subtests. Issues of standardization, such as eliminating words that could be read with different pronunciations in spelling tests, came to the fore, as did tests' *trustworthiness*—a term that, at that time, encompassed what is currently meant by *reliability* and *validity* (DuBois, 1970).

OTHER DEVELOPMENTS IN PSYCHOLOGICAL TESTING

The successes achieved with the Binet and Army tests proved their worth in helping to make decisions about people. This soon led to efforts to devise instruments to help in different kinds of decisions. Naturally, the settings where antecedents of psychological tests had arisen—schools, clinics, and psychology labs—also gave rise to the new forms and types of modern psychological tests.

A thorough review of the history of testing in the first half of the 20th century is beyond the scope of this work. Nevertheless, a brief summary of the most salient developments is instructive both for its own sake and to illustrate the diversity of the field, even in its early phase.

Standardized Testing in Educational Settings

As the number of people availing themselves of educational opportunities at all levels grew, so did the need for fair, equitable, and uniform measures with which to evaluate students at the beginning, middle, and final stages of the educational process. Two major developments in standardized educational testing in the early part of the 20th century are highlighted in the ensuing paragraphs.

Standardized Achievement Tests

Pioneered by E. L. Thorndike, these measures had been under development since the 1880s, when Joseph Rice began his attempts to study the efficiency of learning in schools. Thorndike's handwriting scale, published in 1910, broke new ground in creating a series of handwriting specimens, ranging from very poor to excellent, against which subjects' performance could be compared. Soon after, standardized tests designed to evaluate arithmetic, reading, and spelling skills would follow, until measures of these and other subjects became a staple of elementary and secondary education. Today, standardized achievement tests are used not only in educational settings, but also in the licensing and certification of professionals who have completed their training. They are also used in other situations, including personnel selection, that require the assessment of mastery of a given field of knowledge.

Scholastic Aptitude Tests

In the 1920s objective examinations, based loosely on the Army Alpha test, began to be used in addition to high school grades for the purpose of making admissions decisions in colleges and universities. This momentous development, which culminated in the creation of the Scholastic Aptitude Test (SAT) in 1926, foreshadowed the arrival of many more instruments that are used to select can-

Rapid Reference 1.4

The Big Test

Nicholas Lemann's (1999) book *The Big Test: The Secret History of the American Meritocracy* uses college admissions testing programs, specifically the SAT, to illustrate the intended and unintended consequences that such testing programs can have for society. The large-scale use of standardized test scores for deciding on admissions into leading institutions of higher education was pioneered by James Bryant Conant, president of Harvard University, and Henry Chauncey, the first president of the Educational Testing Service (ETS), in the 1940s and 1950s. Their goal was to change the process whereby access to these institutions—and to the positions of power that usually accrue to those who attend them—is gained from one based on wealth and social class to one based mainly on ability as demonstrated through test scores. Lemann maintains that although this use of testing did open up the doors of higher education to children of the middle and lower socioeconomic classes, it also generated a new meritocratic elite that perpetuates itself across generations and largely excludes the children of underprivileged racial minorities who lack the early educational opportunities needed to succeed on the tests.

didates for graduate and professional schools. Among the best known examples of tests of this type are the Graduate Record Exam (GRE), Medical College Admission Test (MCAT), and Law School Admission Test (LSAT), used by doctoral programs, medical schools, and law schools, respectively. Although each of these tests contains portions specific to the subject matter of its field, they also typically share a common core that emphasizes the verbal, quantitative, and reasoning abilities needed for success in most academic endeavors. Interestingly, although their purpose is different from that of the standardized achievement tests, their content is often similar. Rapid Reference 1.4 presents information about a fascinating account of the history of higher education admissions testing in the United States.

Personnel Testing and Vocational Guidance

The optimal utilization of people's talents is a major goal of society to which psychological testing has been able to contribute in important ways almost from its beginnings. Decisions concerning vocational choice need to be made by individuals at different points in their lives, usually during adolescence and young adulthood but also increasingly at midlife. Decisions concerning the selection and placement of personnel within business, industry, and military organizations

need to be made on an ongoing basis. Some of the main instruments that came into being early and have proved to be particularly helpful in making both of these kinds of decisions are described in the following sections.

Tests of Special Skills and Aptitudes

The success of the Army Alpha test stimulated interest in developing tests to select workers for different occupations. At the same time, applied psychologists had been working out and using a basic set of procedures that would justify the use of tests in occupational selection. Basically, the procedures involved (a) identifying the skills needed for a given occupational role by means of a *job analysis,* (b) administering tests designed to assess those skills, and (c) correlating the test results with measures of job performance. Using variations of this procedure, from the 1920s on, psychologists were able to develop instruments for selecting trainees in fields as diverse as mechanical work and music. Tests of clerical, spatial, and motor abilities soon followed. The field of personnel selection in industry and the military grew up around these instruments, along with the use of job samples, biographical data, and general intelligence tests of the individual and group types. Many of the same instruments have also been used profitably in identifying the talents of young people seeking vocational guidance.

Multiple Aptitude Batteries

The use of tests of separate abilities in vocational counseling would largely give way in the 1940s to multiple aptitude batteries, developed through the factor analytic techniques pioneered by Spearman and expanded in England and the United States through the 1920s and 1930s. These batteries are groups of tests, linked by a common format and scoring basis, that typically profile the strengths and weaknesses of an individual by providing separate scores on various factors such as verbal, numerical, spatial, logical reasoning, and mechanical abilities, rather than the single global score provided by the Binet and Army test IQs. Multiple aptitude batteries came into being following the widespread realization, through factor analyses of ability test data, that intelligence is not a unitary concept and that human abilities comprise a broad range of separate and relatively independent components or factors.

Measures of Interests

Just as tests of special skills and aptitudes arose in industry and later found some use in vocational counseling, measures of interests originated for the purpose of vocational guidance and later found some use in personnel selection. Truman L. Kelley, in 1914, produced a simple Interest Test, possibly the first interest inventory ever, with items concerning preferences for reading materials and leisure ac-

tivities as well as some involving knowledge of words and general information. However, the breakthrough in this particular area of testing took place in 1924, when M. J. Ream developed an empirical key that differentiated the responses of successful and unsuccessful salesmen on the Carnegie Interest Inventory developed by Yoakum and his students at the Carnegie Institute of Technology in 1921 (DuBois, 1970). This event marked the beginning of a technique known as *empirical criterion keying,* which, after refinements such as cross-validation procedures and extensions to other occupations, would be used in the Strong Vocational Interest Blank (SVIB), first published in 1927, and in other types of inventories as well. The current version of the SVIB—called the Strong Interest Inventory® (SII)—is one of the most widely used interest inventories and has been joined by many more instruments of this type.

Clinical Testing

By the start of the 20th century the field of psychiatry had embarked on more systematic ways of classifying and studying psychopathology. These advances provided the impetus for the development of instruments that would help diagnose psychiatric problems. The main examples of this type of tools are discussed here.

Personality Inventories

The first device of this kind was the Woodworth Personal Data Sheet (P-D Sheet), a questionnaire developed during World War I to screen recruits who might suffer from mental illnesses. It consisted of 116 statements regarding feelings, attitudes, and behaviors obviously indicative of psychopathology to which the respondent answered simply yes or no. Although the P-D Sheet showed some promise, World War I ended before it was placed into operational use. After the war there was a period of experimentation with other, less obvious, kinds of items and with scales designed to assess neuroticism, personality traits—such as introversion and extraversion—and values. Innovations in the presentation of items aimed at reducing the influence of social desirability, like the forced-choice technique introduced in the Allport-Vernon Study of Values in 1931, came into being. However, the most successful personality inventory of that era, and one which still survives today, was the Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1940). The MMPI combined items from the P-D Sheet and other inventories, but used the empirical criterion keying technique pioneered with the SVIB. This technique resulted in a less transparent instrument on which respondents could not dissemble as easily because many of the items had no obvious reference to psychopathological tendencies.

Since the 1940s, personality inventories have flourished. Many refinements have been introduced in their construction, including the use of theoretical perspectives—such as Henry Murray's (1938) system of needs—and internal consistency methods of selecting items. Furthermore, factor analysis, which had been so crucial to the study and differentiation of abilities, also began to be used in personality inventory development. In the 1930s, J. P. Guilford pioneered the use of factor analysis to group items into homogeneous scales while, in the 1940s, R. B. Cattell applied the technique to try to identify the personality traits that are most pivotal and, therefore, worthy of investigation and assessment. Currently, factor analysis plays an integral role in most facets of test theory and test construction.

Projective Techniques

Although personality inventories had some success, mental health professionals working with psychiatric populations felt a need for additional help in diagnosing and treating mental illness. In the 1920s, a new genre of tools for the assessment of personality and psychopathology emerged. These instruments, known as *projective techniques,* had their roots in the free association methods pioneered by Galton and used clinically by Kraepelin, Jung, and Freud. In 1921, a Swiss psychiatrist named Hermann Rorschach published a test consisting of ten inkblots to be presented for interpretation, one at a time, to the examinee. The key to the success of this first formal projective technique was that it provided a standardized method for obtaining and interpreting subjects' responses to the inkblot cards, responses that—by and large—reflect the subject's unique modes of perceiving and relating to the world. Rorschach's test was taken up by several American psychologists and propagated in various universities and clinics in the United States after his untimely death in 1922. The Rorschach technique, along with other pictorial, verbal, and drawing instruments, like the Thematic Apperception Test, sentence completion tests, and human figure drawings provided a whole new repertoire of tools—more subtle and incisive than the questionnaires—with which clinicians could investigate aspects of personality that test takers themselves may have been unable or unwilling to reveal. Though there is much controversy about their validity, primarily because they often rely on qualitative interpretations as much as or more than on numerical scores, projective techniques are still a significant part of the toolkit of many clinicians (Viglione & Rivera, 2003).

Neuropsychological Tests

The role of brain dysfunction in emotional, cognitive, and behavioral disorders has been increasingly recognized throughout the past century. However, the major impetus for the scientific and clinical study of brain-behavior relationships,

which is the subject of *neuropsychology,* came from Kurt Goldstein's investigations of the difficulties he observed in soldiers who had sustained brain injuries during World War I. Often these soldiers showed a pattern of deficits involving problems with abstract thinking, memory, as well as the planning and execution of relatively simple tasks, all of which came to be known under the rubric of *organicity,* which was used as a synonym for brain damage. Over several decades, a number of instruments meant to detect organicity, and distinguish it from other psychiatric disorders, came into being. Many of these were variations of the performance—as opposed to verbal—tests that had been developed to assess general intellectual ability in individuals who could not be examined in English or who had hearing or speech impairments. These tests involved materials like form boards, jigsaw puzzles, and blocks as well as paper-and-pencil tasks such as mazes and drawings. A great deal has been learned about the brain and its functioning in the past few decades and much of the initial thinking in neuropsychological assessment has had to be revised based on new information. Brain damage is no longer viewed as an all-or-none condition of organicity with a common set of symptoms, but rather as a huge range of possible disorders resulting from the interaction of specific genetic and environmental factors in each individual case. Nevertheless, the field of neuropsychological assessment has continued to grow in the number and types of instruments available and has contributed both to the clinical and scientific understanding of the many and varied relationships between brain functioning and cognition, emotions, and behaviors (Lezak, 1995).

CURRENT USES OF PSYCHOLOGICAL TESTS

Present-day testing is, on the whole, more methodologically sophisticated and better informed than at any time in the past. The current uses of tests, which take place in a wide variety of contexts, may be classified into three categories: (a) decision-making, (b) psychological research, and (c) self-understanding and personal development. As can be gleaned from this list, presented in Rapid Reference 1.5, the

Rapid Reference 1.5

Current Uses of Psychological Tests

- The first and foremost use of tests is in the pragmatic process of *making decisions about people,* either as individuals or as groups.
- The second use of tests in terms of frequency and longevity is in scientific *research on psychological phenomena* and individual differences.
- The most recent, and least developed, use of tests is in the therapeutic process of *promoting self-understanding and psychological adjustment.*

three kinds of uses differ vastly in their impact and in many other respects, and the first one of them is by far the most visible to the public.

Decision Making

The primary use of psychological tests is as decision-making tools. This particular application of testing invariably involves value judgments on the part of one or more decision makers who need to determine the bases upon which to select, place, classify, diagnose, or otherwise deal with individuals, groups, organizations, or programs. Naturally, this use of testing is often fraught with controversy since it often results in consequences that are unfavorable for one or more parties. In many situations in which tests are used to make decisions and people disagree with the decisions made, the use of tests itself is attacked regardless of whether or not it was appropriate.

When tests are used for making significant decisions about individuals or programs, testing should be merely a part of a thorough and well-planned decision-making strategy that takes into account the particular context in which the decisions are made, the limitations of the tests, and other sources of data in addition to tests. Unfortunately, very often—for reasons of expediency, carelessness, or lack of information—tests are made to bear the responsibility for flawed decision-making processes that place too much weight on test results and neglect other pertinent information. A number of decisions made by educational, governmental, or corporate institutions on a routine basis, usually involving the simultaneous evaluation of several people at once, have been and still are made in this fashion. Although they carry important consequences—such as employment, admission to colleges or professional schools, graduation, or licensure to practice a profession—for the individuals involved, decisions are based almost exclusively on test scores. This practice, a legacy of the way in which testing originated, is one that testing professionals, as well as some government agencies, are trying to change. One of several important steps in this direction is the publication of a resource guide for educators and policymakers on the use of tests as part of high-stakes decision making for students (U.S. Department of Education, Office for Civil Rights, 2000).

Psychological Research

Tests are often used in research in the fields of differential, developmental, abnormal, educational, social, and vocational psychology, among others. They provide a well-recognized method of studying the nature, development, and inter-

relationships of cognitive, affective, and behavioral traits. In fact, although a number of tests that originated in the course of psychological investigations have become commercially available, many more instruments remain archived in dissertations, journals, and various compendiums of experimental measures discussed in Sources of Information about Tests at the end of this chapter. Because there are seldom any immediate practical consequences attendant to the use of tests in research, their use in this context is less contentious than when they are used in decision making about individuals, groups, organizations, or programs.

Self-Understanding and Personal Development

Most humanistic psychologists and counselors have traditionally perceived the field of testing, often justifiably, as overemphasizing the labeling and categorization of individuals in terms of rigid numerical criteria. Starting in the 1970s, a few of them, notably Constance Fischer (1985/1994), began to use tests and other assessment tools in an individualized manner, consonant with humanistic and existential-phenomenological principles. This practice, which views testing as a way to provide clients with information to promote self-understanding and positive growth, has evolved into the *therapeutic model of assessment* espoused by Finn and Tonsager (1997). Obviously, the most pertinent application of this model is in counseling and psychotherapeutic settings in which the client is the main and only user of test results.

PSYCHOLOGICAL ASSESSMENT VERSUS PSYCHOLOGICAL TESTING

For reasons that are mostly related to the marketing of tests, some test authors and publishers have begun to use the word *assessment* in the titles of their tests. Thus, in the mind of the general public the terms *assessment* and *testing* are often seen as synonymous. This is an unfortunate development. The distinction between these terms is one that many people in the field believe is worth preserving, and one that the general public, as potential assessment clients or consumers of tests, should be aware of as well.

The use of tests for making decisions about a person, a group, or a program should always take place within the context of *psychological assessment*. This process can occur in

DON'T FORGET

- *Tests* and *assessments* are NOT synonymous.
- Tests are among the tools used in the process of assessment.

health care, counseling, or forensic settings, as well as in educational and employment settings. Psychological assessment is a flexible, not standardized, *process* aimed at reaching a defensible determination concerning one or more psychological issues or questions, through the collection, evaluation, and analysis of data appropriate to the purpose at hand (Maloney & Ward, 1976).

Steps in the Assessment Process

The first and most important step in psychological assessment is to identify its goals as clearly and realistically as possible. Without clearly defined objectives that are agreed upon by the assessor and the person requesting the assessment, the process is not likely to be satisfactory. In most instances, the process of assessment ends with a verbal or written report, communicating the conclusions that have been reached to the persons who requested the assessment, in a comprehensible and useful manner. In between these two points, the professional conducting the assessment, usually a psychologist or a counselor, will need to employ her or his expertise at every step. These steps involve the appropriate selection of instruments to be used in gathering data, their careful administration, scoring, interpretation, and—most important of all—the judicious use of the data collected to make inferences about the question at hand. This last step goes beyond psychometric expertise and requires a knowledge of the field to which the question refers, such as health care, educational placement, psychopathology, organizational behavior, or criminology, among others. Examples of issues amenable to investigation through psychological assessment include

- *diagnostic questions,* such as differentiating between depression and dementia;
- *making predictions,* such as estimating the likelihood of suicidal or homicidal behaviors; and
- *evaluative judgments,* such as those involved in child custody decisions or in assessing the effectiveness of programs or interventions.

None of these complex issues can be resolved by means of test scores alone because the same test score can have different meanings depending on the examinee and the context in which it was obtained. Furthermore, no single test score or set of scores can capture all the aspects that need to be considered in resolving such issues.

Psychological tests may be key components in psychological assessment, but the two differ fundamentally in important ways. Rapid Reference 1.6 lists several dimensions that differentiate psychological testing and assessment. Even though

Rapid Reference 1.6

Typical Differences between Psychological Testing and Assessment

Basis	Psychological Testing	Psychological Assessment
Degree of complexity	Simpler; involves one uniform procedure, frequently unidimensional.	More complex; each assessment involves various procedures (interviewing, observation, testing, etc.) and dimensions.
Duration	Shorter, lasting from a few minutes to a few hours.	Longer, lasting from a few hours to a few days or more.
Sources of data	One person, the test taker.	Often collateral sources, such as relatives or teachers, are used in addition to the subject of the assessment.
Focus	How one person or group compares with others (*nomothetic*).	The uniqueness of a given individual, group, or situation (*idiographic*).
Qualifications for use	Knowledge of tests and testing procedures.	Knowledge of testing and other assessment methods as well as of the area assessed (e.g., psychiatric disorders, job requirements).
Procedural basis	Objectivity required; quantification is critical.	Subjectivity, in the form of clinical judgment, required; quantification rarely possible.
Cost	Inexpensive, especially when testing is done in groups.	Very expensive; requires intensive use of highly qualified professionals.
Purpose	Obtaining data for use in making decisions.	Arriving at a decision concerning the referral question or problem.
Degree of structure	Highly structured.	Entails both structured and unstructured aspects.
Evaluation of results	Relatively simple investigation of reliability and validity based on group results.	Very difficult due to variability of methods, assessors, nature of presenting questions, etc.

there is little question about the general superiority of assessment over testing with regard to comprehensiveness and utility, the greater complexity of the assessment process makes its results far more difficult to evaluate than those of testing. Nevertheless, in recent years, evidence of the efficacy of assessment, at least in the realm of health care delivery, has begun to be assembled (Eisman et al., 2000; Kubiszyn et al., 2000; Meyer et al., 2001).

TEST USER QUALIFICATIONS

As the number of tests has continued to grow and their uses have expanded, not only in the United States but around the world, the question of test misuse has become of increasing concern for the public, the government, and various professions. Psychology, which is the profession from which tests arose and the one with which they are most distinctly associated, has taken the lead in trying to combat their misuse. The *Testing Standards* promulgated by the APA and other professional organizations (AERA, APA, & NCME, 1999) are a major vehicle to this end. The APA also addresses issues related to testing and assessment in its ethical principles and code of conduct (APA, 2002), as do other professional associations (e.g., American Counseling Association, 1995; National Association of School Psychologists, 2000).

Although the technical qualities of a number of tests are far from ideal and can contribute to problems in their use, it is generally conceded that the primary reason for test misuse lies in the insufficient knowledge or competence on the part of many test users. Tests may appear relatively simple and straightforward to potential users who are unaware of the cautions that must be exercised in their application. Because of this, in the past few decades, professional associations in the United States and elsewhere have been developing documents that outline more clearly and specifically than ever before the skills and knowledge base required for competent test use (American Association for Counseling and Development, 1988; Eyde, Moreland, Robertson, Primoff, & Most, 1988; International Test Commission, 2000; Joint Committee on Testing Practices, 1988).

One of the clearest expositions of these requirements is in a report prepared over the course of five years by the APA Task Force on Test User Qualifications (APA, 2000). This report outlines (a) the core knowledge and skills essential to those who use tests to make decisions or formulate policies that affect the lives of test takers, and (b) the expertise that test users in the specific contexts of employment, education, career counseling, health care, and forensic work must possess. Core or generic knowledge and skills in psychometrics, statistics, test selection, administration, scoring, reporting, and safeguarding are considered relevant

to all test users. Additional knowledge and supervised experience required for the use of tests in the various contexts and with diverse groups of test takers are also outlined in the report, as are the variety of uses of tests in classification, description, prediction, intervention planning, and tracking in each context.

Another aspect of testing that has contributed to test misuse over the decades is the relative ease with which test instruments can be obtained by people who are not qualified to use them. To some extent, the availability of tests is a function of the freedom with which information flows in democratic societies like the United States, especially in the era of the World Wide Web. Another reason for this problem—alluded to earlier in this chapter—is the fact that many tests are commercial products. As a result, some test publishers have been willing to sell tests to persons or institutions without using adequate safeguards to ascertain whether they possess the proper credentials. At one point, during the 1950s and 1960s, the *Testing Standards* included a three-tiered system for classifying tests in terms of the qualifications needed for their use (APA, 1966, pp. 10–11). This system, which labeled tests as Level A, B, or C depending on the training required to use them, was easily circumvented by individuals in schools, government agencies, and businesses. Although many test publishers still use the system, the *Testing Standards* no longer do. Rapid Reference 1.7 outlines the elements typically included in a three-tiered classification system of test user qualifications.

In 1992, a number of the publishers of tests and providers of assessment services established the Association of Test Publishers (ATP). This nonprofit organization tries to uphold a high level of professionalism and ethics in the testing enterprise. One way in which they monitor the distribution of tests is by requiring some documentation attesting to a minimum level of training from those who would purchase their products.

Qualification forms for test purchase are now included in the catalogs of all reputable test publishers. No matter how sincere publishers may be in their efforts to preserve the security of test materials and to prevent their misuse, the effectiveness of these efforts is by necessity limited. Not only is it not feasible to verify the qualifications that purchasers claim on the forms they submit, but in addition no formal set of qualifications—whether by education or by licensure—can ensure that an individual is competent to use a particular test properly in a given situation (see Chapter 7).

SOURCES OF INFORMATION ABOUT TESTS

In psychological testing, as in every other human endeavor, the Internet has created an inexhaustible supply of information. Thus, alongside the print references

Rapid Reference 1.7

Test User Qualification Levels

All reputable test publishers require test purchasers to complete a form specifying the credentials that qualify them to use the testing materials they wish to buy and certifying that the materials will be used in accordance with all applicable ethical and legal guidelines. Although the number of levels and the specific credentials required at each level differ among publishers, their qualification criteria are typically organized into at least three tiers, based roughly on a categorization of tests and training requirements originally outlined by the American Psychological Association (APA; 1953, 1954).

	Lowest Tier (Level A)	Intermediate Tier (Level B)	Highest Tier (Level C)
Type of instruments to which this level applies	A limited range of instruments, such as educational achievement tests, that can be administered, scored, and interpreted without specialized training, by following the instructions in their manuals.	Tools that call for some specialized training in test construction and use and in the area in which the instruments will be applied, such as aptitude tests and personality inventories applicable to normal populations.	Instruments that require extensive familiarity with testing and assessment principles, as well as with the psychological fields to which the instruments pertain, such as individual intelligence tests and projective techniques.
Kinds of credentials or requirements necessary to purchase materials at this level	Some publishers do not require any credentials to purchase tests at this level. Others may require a bachelor's degree in an appropriate field or that orders for materials be placed through an agency or institution, or both.	Test purchasers usually must have either a Master's-level degree in psychology (or in a related field), or course work in testing and assessment commensurate with the requirements for using the instruments in question.	Test purchasers must have the kind of advanced training and supervised experience that is acquired in the course of obtaining a doctoral degree, or professional licensure in a field pertinent to the intended use of the instruments, or both.

that the field has traditionally had, there now is a large number of on-line and electronic media resources that are easily accessible.

Internet Resources

For the person who seeks information about psychological tests, a good starting point is the Testing and Assessment section of the APA's Web site (http://www.apa.org). Within this section, among other things, there is an excellent article on "FAQ/Finding Information About Psychological Tests" (APA, 2003) that provides guidance on how to locate published and unpublished tests as well as important documents relevant to psychological testing. *Published tests* are commercially available through a test publisher, although they sometimes go out of print as books do. *Unpublished tests* have to be obtained directly from the individual investigator who created them, unless they appear in the periodical literature or in specialized directories (discussed shortly).

Two other great entry points on the Internet, for those who seek information about a specific test, are (a) the Buros Institute of Mental Measurements (BI) Test Reviews Online Web page (http://www.unl.edu/buros), which offers free information on nearly 4,000 commercially available tests as well as more than 2,000 test reviews that can be purchased and displayed online; and (b) the Educational Testing Service (ETS) Test Collection database (at http://www.ets.org/testcoll/index.html), which is the largest of its kind in the world. In addition, the Educational Resources Information Center (ERIC) system Web site (http://eric.ed.gov)—funded by the U.S. Department of Education—contains a wealth of materials related to psychological testing.

Another way to obtain information about both published and unpublished tests online is through the electronic indexes of the periodical literature in psychology, education, or business. The PsycINFO database of the APA, available through many libraries or by subscription, provides an entry point at which to use the name of a test to find bibliographic references, abstracts, and even full text of articles about it. In addition to exact titles, PsycINFO and other

DON'T FORGET

One of the most basic distinctions among tests concerns whether they are published.

- *Published tests* are commercially available through test publishers.
- *Unpublished tests* must be obtained from the individual investigator who developed them, from special directories of unpublished measures, or from the periodical literature.

DON'T FORGET

Appendix A lists all of the commercially available, published tests and psychological assessment instruments mentioned throughout this book, along with codes identifying their publishers.

Appendix B provides current Internet addresses for the publishers listed in Appendix A. More detailed information on test publishers, including street addresses and telephone numbers, is available in the latest edition of *Tests in Print* (Murphy, Plake, Impara, & Spies, 2002).

databases also can be searched by subjects, keywords, and authors, which makes them especially useful when only partial information is available.

Once a test is located through any of these resources, one can usually also determine whether it is published and how it can be obtained. If the test is published, it may be ordered from the company that publishes it by those who meet the qualifications to use it. Ordering information is available in the publishers' catalogs, many of which are now available online as well as in printed form. The ATP Web site (http://www.testpublishers.org) has links to many test publishers and providers of assessment services. Internet addresses for all of the organizations mentioned in this section, and other important sources of information on tests, can be found in Rapid Reference 1.8.

Rapid Reference 1.8

Internet Sources of Information on Psychological Tests

Organization (Acronym)	Website
American Educational Research Association (AERA)	http://www.aera.net
American Psychological Association (APA)	http://www.apa.org
Association of Test Publishers (ATP)	http://www.testpublishers.org
Buros Institute of Mental Measurements (BI)	http://www.unl.edu/buros
Educational Resources Information Center (ERIC)	http://eric.ed.gov
Educational Testing Service (ETS)	http://www.ets.org/testcoll/index.html
International Test Commission (ITC)	http://www.intestcom.org
National Council on Measurement in Education (NCME)	http://www.ncme.org

Print Resources

Published Tests

As far as commercially available, published tests are concerned, the most important sources of information stem from the Buros Institute of Mental Measurements (BI) in Lincoln, Nebraska. In particular, the BI (http://www.unl.edu/buros) produces two series of volumes that can guide users to almost every published test available in the United States. One of these is the *Tests in Print* (*TIP*) series and the other is the *Mental Measurements Yearbook* (*MMY*) series. *Tests in Print* is a comprehensive bibliography of all tests that are commercially available at the time a given volume of the series is published. Each entry has the test title, acronym, author, publisher, publication date, and other basic information about the test as well as cross-references to the reviews of the test in all the *MMY*s available at that point. In addition, the *TIP* series contains an extremely useful classified index of tests that are in print, as well as indexes of test scores, publishers, acronyms, and names of authors and reviewers. The *MMY* series, in turn, goes back to 1938, when the late Oscar Buros published the first yearbook to assist test users by providing evaluative test reviews written by qualified and independent professionals. Although the *MMY*s are still published in book form, their entries and reviews are also available online and in other electronic media. The Buros Institute also publishes many other test-related materials.

PRO-ED (http://www.proedinc.com) is the publisher of *Tests,* a series of encyclopedic volumes listing short descriptions of instruments in psychology, education, and business. The *Test Critiques* series, dating back to 1984, is the companion to *Tests.* Each volume in this series contains test reviews and cumulative indexes to all its previous volumes.

Unpublished Tests

The goal of behavioral scientists who use psychological tests is to investigate psychological constructs as well as individual and group differences. Many existing tests are used exclusively for scientific research and are not commercially available. These tests are referred to as *unpublished* measures because they cannot be purchased; conditions for their use are typically established by the authors of each instrument and most often require a letter requesting permission to use them. Information about unpublished tests—and often the instruments themselves—is available in the periodical literature in psychology (e.g., through PsycINFO online) and in various directories (e.g., Goldman, Mitchell, & Egelson, 1997; Robinson, Shaver, & Wrightsman, 1991). The previously mentioned article "FAQ/Finding Information About Psychological Tests" (APA, 2003) lists several print and electronic resources for information on unpublished tests.

TEST YOURSELF

1. **Which of the following is *not* an essential element of psychological testing?**
   (a) Systematic procedures
   (b) The use of empirically derived standards
   (c) Preestablished rules for scoring
   (d) Sampling behavior from affective domains
2. **The single most important source of criteria for evaluating tests, testing practices, and the effects of test use can be found in the**
   (a) *Ethical Principles of Psychologists and Code of Conduct.*
   (b) *Standards for Educational and Psychological Testing.*
   (c) *Diagnostic and Statistical Manual of Mental Disorders.*
   (d) *Report of the Task Force on Test User Qualifications.*
3. **The earliest antecedents of modern testing for personnel selection date back to**
   (a) China, B.C.E.
   (b) ancient Greece.
   (c) the Inca empire.
   (d) Medieval Europe.
4. **Evaluating psychological tests is *least* problematic**
   (a) prior to their being placed into use.
   (b) once they have been placed into use.
5. **Compared to the other areas listed, the development of criteria or bases for decision making has been substantially slower in the context of**
   (a) educational assessment.
   (b) occupational assessment.
   (c) clinical assessment.
6. **Credit for devising the first successful psychological test in the modern era is usually given to**
   (a) Francis Galton.
   (b) Alfred Binet.
   (c) James McKeen Cattell.
   (d) Wilhelm Wundt.

7. The true ratio IQ or intelligence quotient was derived by

(a) adding the mental age (MA) and the chronological age (CA) of the test taker.

(b) subtracting the CA from the MA and multiplying the result by 100.

(c) dividing the CA by the MA and multiplying the result by 100.

(d) dividing the MA by the CA and multiplying the result by 100.

8. The primary purpose for which psychological tests are currently used is

(a) psychological research.

(b) educational research.

(c) decision making.

(d) self-understanding and personal development.

9. Compared to psychological testing, psychological assessment is generally

(a) simpler.

(b) more structured.

(c) more expensive.

(d) more objective.

10. Which of the following would be the best source of information on a test that is *not* commercially available?

(a) *Mental Measurements Yearbooks*

(b) *Test Critiques*

(c) *Tests in Print*

(d) PsycINFO

Answers: 1. d; 2. b; 3. a; 4. a; 5. c; 6. b; 7. d; 8. c; 9. c; 10. d.