Psychological Testing

Principles, Applications, and Issues

Seventh Edition

Robert M. Kaplan
Dennis P. Saccuzzo

Psychological Testing: Principles, Applications, and Issues, Seventh Edition

Robert M. Kaplan, Dennis P. Saccuzzo

Editor: Jaime Perkins

Editorial Assistant: Wilson Co

Technology Project Manager: Amy Cohen

Marketing Manager: Kim Russell

Marketing Assistant: Molly Felz

Marketing Communications Manager: Talia Wise

Project Manager, Editorial Production: Charlene M. Carpentier

Creative Director: Rob Hugel

Art Director: Vernon Boes

Print Buyer: Linda Hsu

Permissions Editor: Bob Kauser

Production Service: Newgen–Austin

Text Designer: Lisa Henry

Photo Researcher: Darren Wright

Copy Editor: Mary Ann Grobbel

Cover Designer: Larry Didona

Cover Image: "Geometric shapes below clouds" ©Pete Turner/The Image Bank/Getty Images

Compositor: Newgen

Library of Congress Control Number: 2008927883
Student Edition:
ISBN-13: 978-0-495-09555-2
ISBN-10: 0-495-09555-9

Wadsworth
10 Davis Drive
Belmont, CA 94002-3098
USA

Cengage Learning is a leading provider of customized learning solutions with office locations around the globe, including Singapore, the United Kingdom, Australia, Mexico, Brazil, and Japan. Locate your local office at **international.cengage.com/region.**

Cengage Learning products are represented in Canada by Nelson Education, Ltd.

For your course and learning solutions, visit **academic.cengage.com.**

Purchase any of our products at your local college store or at our preferred online store **www.ichapters.com.**

Printed in the United States of America
1 2 3 4 5 6 7 12 11 10 09 08

CHAPTER 1

Introduction

LEARNING OBJECTIVES

When you have completed this chapter, you should be able to:

- Define the basic terms pertaining to psychological and educational tests
- Distinguish between an individual test and a group test
- Define the terms *achievement, aptitude,* and *intelligence* and identify a concept that can encompass all three terms
- Distinguish between ability tests and personality tests
- Define the term *structured personality test*
- Explain how structured personality tests differ from projective personality tests
- Explain what a normative or standardization sample is and why such a sample is important
- Identify the major developments in the history of psychological testing
- Explain the relevance of psychological tests in contemporary society

You are sitting at a table. You have just been fingerprinted and have shown a picture ID. You look around and see 40 nervous people. A stern-looking test proctor with a stopwatch passes out booklets. You are warned not to open the booklet until told to do so; you face possible disciplinary action if you disobey. This is not a nightmare or some futuristic fantasy—this is real.

Finally, after what seems like an eternity, you are told to open your booklet to page 3 and begin working. Your mouth is dry; your palms are soaking wet. You open to page 3. You have 10 minutes to solve a five-part problem based on the following information.[1]

> A car drives into the center ring of a circus and exactly eight clowns—Q, R, S, T, V, W, Y, and Z—get out of the car, one clown at a time. The order in which the clowns get out of the car is consistent with the following conditions:
>
> V gets out at some time before both Y and Q.
> Q gets out at some time after Z.
> T gets out at some time before V but at some time after R.
> S gets out at some time after V.
> R gets out at some time before W.
>
> *Question 1*. If Q is the fifth clown to get out of the car, then each of the following could be true *except*:
>
> Z is the first clown to get out of the car.
> T is the second clown to get out of the car.
> V is the third clown to get out of the car.
> W is the fourth clown to get out of the car.
> Y is the sixth clown to get out of the car.

Not quite sure how to proceed, you look at the next question.

> *Question 2*. If R is the second clown to get out of the car, which of the following must be true?
>
> S gets out of the car at some time before T does.
> T gets out of the car at some time before W does.
> W gets out of the car at some time before V does.
> Y gets out of the car at some time before Q does.
> Z gets out of the car at some time before W does.

Your heart beats a little faster and your mind starts to freeze up like an overloaded computer with too little working memory. You glance at your watch and notice that 2 minutes have elapsed and you still don't have your bearings. The person sitting next to you looks a bit faint. Another three rows up someone storms up to the test proctor and complains frantically that he cannot do this type of problem. While the proctor struggles to calm this person down, another makes a mad dash for the restroom.

Welcome to the world of competitive, "high stakes," standardized psychological tests in the 21st century. The questions you just faced were actual problems from

---

[1]Used by permission from the Law School Admission Test, October 2002. Answer to Question 1 is D; answer to Question 2 is E.

a past version of the LSAT—the Law School Admission Test. Whether or not a student is admitted into law school in the United States is almost entirely determined by that person's score on the LSAT and undergraduate college grade point average. Thus, one's future can depend to a tremendous extent on a single score from a single test given in a tension-packed morning or afternoon. Despite efforts to improve tests like the LSAT to increase diversity (Pashley, Thornton, & Duffy, 2005), standardized tests tend to disadvantage women and ethnic minorities (Sackett, Schmitt, Ellingson, & Kabin, 2001). Similar problems appear on the GRE—the Graduate Record Exam, a test that plays a major role in determining who gets to study at the graduate level in the United States. (Later in this book we discuss how to prepare for such tests and what their significance, or predictive validity, is.)

Tests such as the LSAT and GRE are the most difficult modern psychological tests. The scenes we've described are real; some careers do ride on a single test. Perhaps you have already taken the GRE or LSAT. Or perhaps you have not graduated yet but are thinking about applying for an advanced degree or professional program and will soon be facing the GRE, LSAT, or MCAT (Medical College Admission Test). Clearly, it will help you to have a basic understanding of the multitude of psychological tests people are asked to take throughout their lives.

From our birth, tests have a major influence on our lives. When the pediatrician strokes the palms of our hands and the soles of our feet, he or she is performing a test. When we enter school, tests decide whether we pass or fail classes. Testing may determine if we need special education. In the United States and many industrialized countries competence tests determine if students will graduate from high school (Carnoy, 2005; Hursh, 2005). More tests determine which college we may attend. And, of course, when we get into college we face still more tests.

After graduation, those who choose to avoid tests such as the GRE may need to take tests to determine where they will work. In the modern world, a large part of everyone's life and success depends on test results. Indeed, tests even have international significance.

For example, 15-year-old children in 32 nations were given problems such as the following from the Organization for Economic Co-operation and Development (OECD) and the Programme for International Student Assessment (PISA) (Schleicher & Tamassia, 2000):

> A result of global warming is that ice of some glaciers is melting.
>
> Twelve years after the ice disappears, tiny plants, called lichen, start to grow on the rocks. Each lichen grows approximately in the shape of a circle.
>
> The relationship between the diameter of the circles and the age of the lichen can be approximated with the formula: $d = 7.0 \times$ the square root of $(t - 12)$ for any t less than or equal to 12, where d represents the diameter of the lichen in millimeters, and t represents the number of years after the ice has disappeared.
>
> Calculate the diameter of the lichen 16 years after the ice disappeared. The complete and correct answer is:

$$d = 7.0 \times \text{the square root of } (16 - 12 \text{ mm})$$

$$d = 7.0 \times \text{the square root of } 4 \text{ mm}$$

$$d = 14 \text{ mm}$$

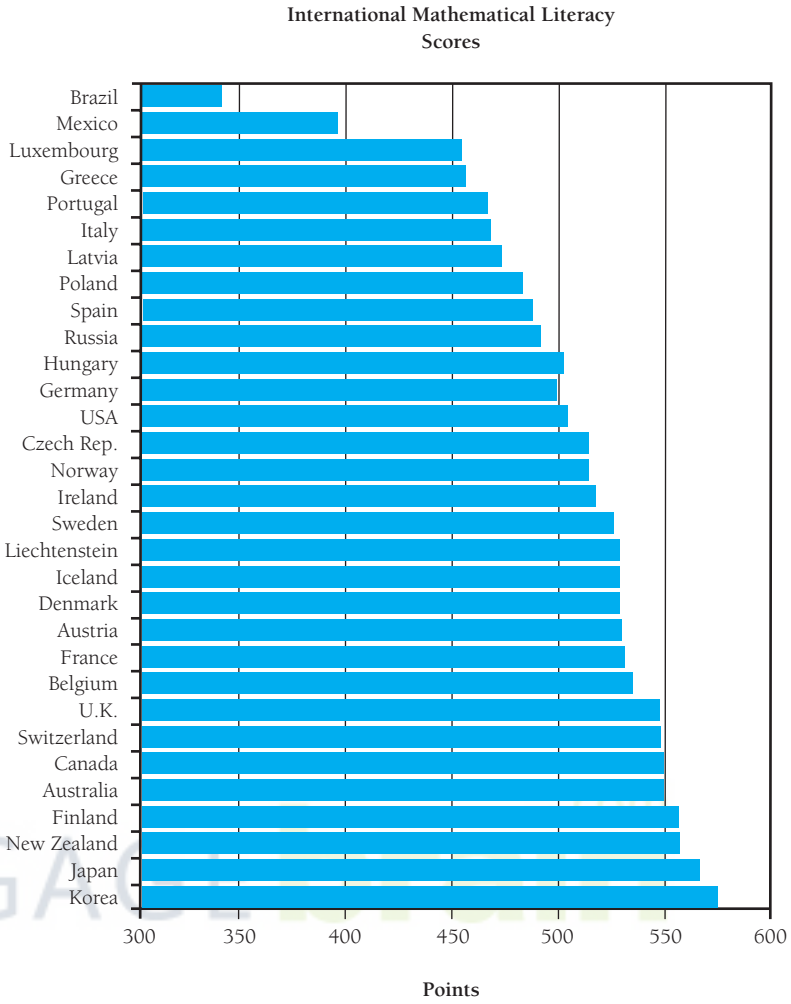


FIGURE **1.1** Approximate average scores of 15-year-old students on the OECD mathematical literacy test.

(Statistics used by permission of the OECD and PISA. Figure courtesy of W. J. Koen.)

Eighteen countries ranked above the United States in the percentage of 15-year-olds who had mastered such concepts (see Figure 1.1).

The results were similar for an OECD science literacy test (see Figure 1.2), which had questions such as the following:

> A bus is moving along a straight stretch of road. The bus driver, named Ray, has a cup of water resting in a holder on the dashboard. Suddenly Ray has to slam on the brakes. What is most likely to happen to the water in the cup immediately after Ray slams on the brakes?
>
> **A.** The water will stay horizontal.
> **B.** The water will spill over side 1.

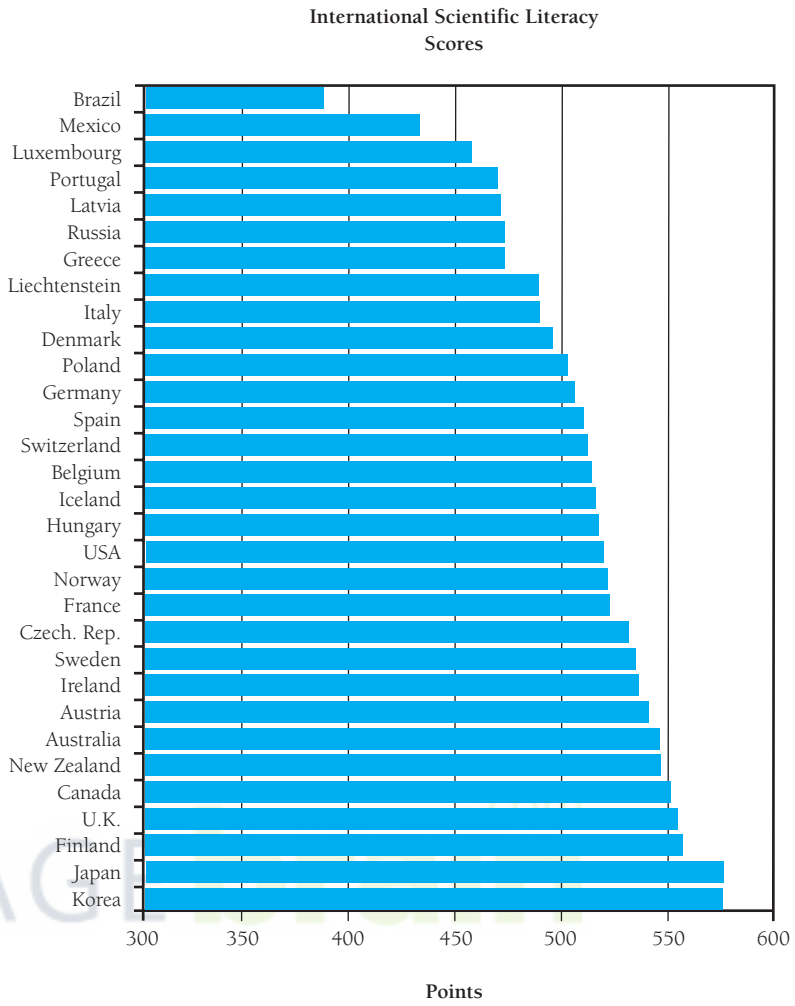


FIGURE **1.2** Approximate average scores of 15-year-old students on the OECD scientific literacy test.

(Statistics used by permission of the OECD and PISA. Figure courtesy of W. J. Koen.)

C. The water will spill over side 2.
D. The water will spill but you cannot tell if it will spill over side 1 or side 2.

The correct answer is C.

How useful are tests such as these? Do they measure anything meaningful? How accurate are they? Such questions concern not only every U.S. citizen but also all members of the highly competitive international community. To answer them, you must understand the principles of psychological testing that you are about to learn.

To answer questions about tests, you must understand the concepts presented in this book, such as reliability, validity, item analysis, and test construction. A full

understanding of these concepts will require careful study and a knowledge of basic statistics, but your efforts will be richly rewarded. When you finish this book, you will be a better consumer of tests.

BASIC CONCEPTS

You are probably already familiar with some of the elementary concepts of psychological testing. For the sake of clarity, however, we shall begin with definitions of the most basic terms so that you will know how they are used in this textbook.

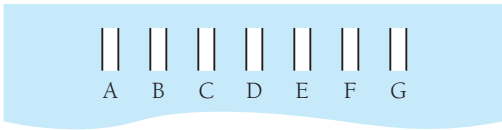
What a Test Is

Everyone has had experience with tests. A **test** is a measurement device or technique used to quantify behavior or aid in the understanding and prediction of behavior. A spelling test, for example, measures how well someone spells or the extent to which someone has learned to spell a specific list of words. At some time during the next few weeks, your instructor will likely want to measure how well you have learned the material in this book. To accomplish this, your instructor may give you a test.

As you well know, the test your instructor gives may not measure your full understanding of the material. This is because a test measures only a sample of behavior, and error is always associated with a sampling process. Test scores are not perfect measures of a behavior or characteristic, but they do add significantly to the prediction process, as you will see.

An **item** is a specific stimulus to which a person responds overtly; this response can be scored or evaluated (for example, classified, graded on a scale, or counted). Because psychological and educational tests are made up of items, the data they produce are explicit and hence subject to scientific inquiry.

In simple terms, items are the specific questions or problems that make up a test. The problems presented at the beginning of this chapter are examples of test items. The overt response would be to fill in or blacken one of the spaces:



A **psychological test** or educational test is a set of items that are designed to measure characteristics of human beings that pertain to behavior. There are many types of behavior. *Overt* behavior is an individual's observable activity. Some psychological tests attempt to measure the extent to which someone might engage in or "emit" a particular overt behavior. Other tests measure how much a person has previously engaged in some overt behavior. Behavior can also be *covert*—that is, it takes place within an individual and cannot be directly observed. For example, your feelings and thoughts are types of covert behavior. Some tests attempt to measure such behavior. Psychological and educational tests thus measure past or current behavior. Some also attempt to predict future behavior, such as success in college or in an advanced degree program.

What does it mean when someone gets 75 items correct on a 100-item test? One thing it means, of course, is that 75% of the items were answered correctly.

In many situations, however, knowing the percentage of correct items a person obtained can be misleading. Consider two extreme examples. In one case, out of 100 students who took the exam, 99 had 90% correct or higher, and 1 had 75% correct. In another case, 99 of the 100 students had scores of 25% or lower, while 1 had 75% correct. The meaning of the scores can change dramatically, depending on how a well-defined sample of individuals scores on a test. In the first case, a score of 75% is poor because it is in the bottom of the distribution; in the second case, 75% is actually a top score. To deal with such problems of interpretation, psychologists make use of **scales**, which relate raw scores on test items to some defined theoretical or empirical distribution. Later in the book you will learn about such distributions.

Scores on tests may be related to traits, which are enduring characteristics or tendencies to respond in a certain manner. "Determination," sometimes seen as "stubbornness," is an example of a trait; "shyness" is another. Test scores may also be related to the state, or the specific condition or status, of an individual. A determined individual after many setbacks may, for instance, be in a weakened state and therefore be less inclined than usual to manifest determination. Tests measure many types of behavior.

Types of Tests

Just as there are many types of behavior, so there are many types of tests. Those that can be given to only one person at a time are known as **individual tests** (see Figure 1.3). The examiner or **test administrator** (the person giving the test) gives the test to only one person at a time, the same way that psychotherapists see only one person at a time. A **group test**, by contrast, can be administered to more than one person at a time by a single examiner, such as when an instructor gives everyone in the class a test at the same time.

One can also categorize tests according to the type of behavior they measure. Ability tests contain items that can be scored in terms of speed, accuracy, or both. On an ability test, the faster or the more accurate your responses, the better your scores on a particular characteristic. The more algebra problems you can correctly solve in a given amount of time, the higher you score in ability to solve such problems.

Historically, experts have distinguished among achievement, aptitude, and intelligence as different types of ability. **Achievement** refers to previous learning. A test that measures or evaluates how many words you can spell correctly is called a *spelling achievement test*. **Aptitude**, by contrast, refers to the potential for learning or acquiring a specific skill. A spelling aptitude test measures how many words you might be able to spell given a certain amount of training, education, and experience. Your musical aptitude refers in part to how well you might be able to learn to play a musical instrument given a certain number of lessons. Traditionally distinguished from achievement and aptitude, **intelligence** refers to a person's general potential to solve problems, adapt to changing circumstances, think abstractly, and profit from experience. When we say a person is "smart," we are usually referring to intelligence. When a father scolds his daughter because she has not done as well in school as she can, he most likely believes that she has not used her intelligence (general potential) to achieve (acquire new knowledge).

The distinctions among achievement, aptitude, and intelligence are not always so cut-and-dried because all three are highly interrelated. Attempts to separate prior learning from potential for learning, for example, have not succeeded. In view of

FIGURE **1.3** An individual test administration.

(Ann Chwatsky/Jeroboam.)

the considerable overlap of achievement, aptitude, and intelligence tests, all three concepts are encompassed by the term **human ability**.

There is a clear-cut distinction between ability tests and personality tests. Whereas ability tests are related to capacity or potential, **personality tests** are related to the overt and covert dispositions of the individual—for example, the tendency of a person to show a particular behavior or response in a given situation. Remaining isolated from others, for instance, does not require any special skill or ability, but some people typically prefer or tend to remain thus isolated. Personality tests measure typical behavior.

There are several types of personality tests. In Chapter 13, you will learn about structured, or objective, personality tests. **Structured personality tests** provide a statement, usually of the "self-report" variety, and require the subject to choose between two or more alternative responses such as "True" or "False" (see Figure 1.4).

In contrast to structured personality tests, projective personality tests are unstructured. In a **projective personality test**, either the stimulus (test materials) or the required response—or both—are ambiguous. For example, in the highly controversial Rorschach test, the stimulus is an inkblot. Furthermore, rather than being asked to choose among alternative responses, as in structured personality tests, the individual is asked to provide a spontaneous response. The inkblot is presented to the subject, who is asked, "What might this be?" Projective tests assume that a person's interpretation of an ambiguous stimulus will reflect his or her unique characteristics (see Chapter 14).

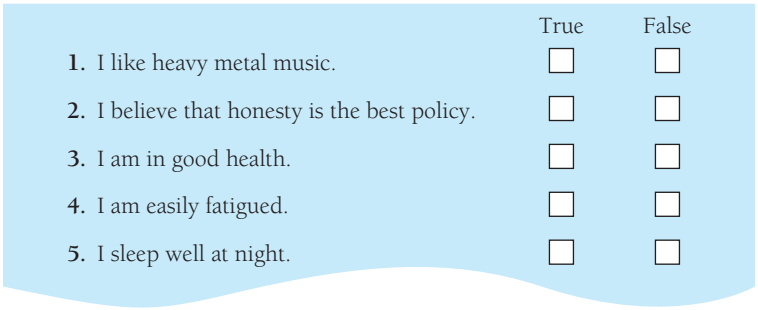


FIGURE **1.4** Self-report test items.

TABLE **1.1**
Types of Tests

I.	**Ability tests:** Measure skills in terms of speed, accuracy, or both.
	A. **Achievement:** Measures previous learning.
	B. **Aptitude:** Measures potential for acquiring a specific skill.
	C. **Intelligence:** Measures potential to solve problems, adapt to changing circumstances, and profit from experience.
II.	**Personality tests:** Measure typical behavior—traits, temperaments, and dispositions.
	A. **Structured (objective):** Provides a self-report statement to which the person responds "True" or "False," "Yes" or "No."
	B. **Projective:** Provides an ambiguous test stimulus; response requirements are unclear.

See Table 1.1 for a brief overview of ability and personality tests.

Psychological testing refers to all the possible uses, applications, and underlying concepts of psychological and educational tests. The main use of these tests, though, is to evaluate individual differences or variations among individuals. Such tests measure individual differences in ability and personality and assume that the differences shown on the test reflect actual differences among individuals. For instance, individuals who score high on an IQ test are assumed to have a higher degree of intelligence than those who obtain low scores. Thus, the most important purpose of testing is to differentiate among those taking the tests. We shall discuss the idea of individual differences later in this chapter.

OVERVIEW OF THE BOOK

This book is divided into three parts: *Principles*, *Applications*, and *Issues*. Together, these parts cover psychological testing from the most basic ideas to the most complex. Basic ideas and events are introduced early and stressed throughout to reinforce

what you have just learned. In covering principles, applications, and issues, we intend to provide not only the *who*'s of psychological testing but also the *how*'s and *why*'s of major developments in the field. We also address an important concern of many students—relevance—by examining the diverse uses of tests and the resulting data.

Principles of Psychological Testing

By *principles of psychological testing* we mean the basic concepts and fundamental ideas that underlie all psychological and educational tests. Chapters 2 and 3 present statistical concepts that provide the foundation for understanding tests. Chapters 4 and 5 cover two of the most fundamental concepts in testing: reliability and validity. **Reliability** refers to the accuracy, dependability, consistency, or repeatability of test results. In more technical terms, reliability refers to the degree to which test scores are free of measurement errors. As you will learn, there are many ways a test can be reliable. For example, test results may be reliable over time, which means that when the same test is given twice within any given time interval, the results tend to be the same or highly similar. **Validity** refers to the meaning and usefulness of test results. More specifically, validity refers to the degree to which a certain inference or interpretation based on a test is appropriate. When one asks the question, "What does this psychological test measure?" one is essentially asking "For what inference is this test valid?"

Another principle of psychological testing concerns how a test is created or constructed. In Chapter 6, we present the principles of test construction. The act of giving a test is known as **test administration**, which is the main topic of Chapter 7. Though some tests are easy to administer, others must be administered in a highly specific way. The final chapter of Part I covers the fundamentals of administering a psychological test.

Applications of Psychological Testing

Part II, on applications, provides a detailed analysis of many of the most popular tests and how they are used or applied. It begins with an overview of the essential terms and concepts that relate to the application of tests. Chapter 8 discusses interviewing techniques. An **interview** is a method of gathering information through verbal interaction, such as direct questions. Not only has the interview traditionally served as a major technique of gathering psychological information in general, but also data from interviews provide an important complement to test results.

Chapters 9 and 10 cover individual tests of human ability. In these chapters, you will learn not only about tests but also about the theories of intelligence that underlie them. In Chapter 11, we cover testing in education with an emphasis on special education. In Chapter 12, we present group tests of human ability. Chapter 13 covers structured personality tests, and Chapter 14 covers projective personality tests. In Chapter 15, we discuss the important role of computers in the testing field. We also consider the influence of cognitive psychology, which today is the most prominent of the various schools of thought within psychology (Kellogg, 2003; Leahy & Dowd, 2002; Weinstein & Way, 2003).

These chapters not only provide descriptive information but also delve into the ideas underlying the various tests. Chapter 16 examines interest tests, which

measure behavior relevant to such factors as occupational preferences. Chapter 17 reviews the relatively new area of medical testing for brain damage and health status. It also covers important recent advancements in developmental neuropsychology. Finally, Chapter 18 covers tests for industrial and organizational psychology and business.

Issues of Psychological Testing

Many social and theoretical issues, such as the controversial topic of racial differences in ability, accompany testing. Part III covers many of these issues. As a compromise between breadth and depth of coverage, we focus on a comprehensive discussion of those issues that have particular importance in the current professional, social, and political environment.

Chapter 19 examines test bias, one of the most volatile issues in the field (Geisinger, 2003; Reynolds & Ramsay, 2003; Ryan & DeMark, 2002). Because psychological tests have been accused of being discriminatory or biased against certain groups, this chapter takes a careful look at both sides of the argument. Because of charges of bias and other problems, psychological testing is increasingly coming under the scrutiny of the law (Phillips, 2002; Saccuzzo, 1999). Chapter 20 examines test bias as related to legal issues and discusses testing and the law. Chapter 21 presents a general overview of other major issues currently shaping the future of psychological testing in the United States with an emphasis on ethics. From our review of the issues, we also speculate on what the future holds for psychological testing.

HISTORICAL PERSPECTIVE

We now briefly provide the historical context of psychological testing. This discussion touches on some of the material presented earlier in this chapter.

Early Antecedents

Most of the major developments in testing have occurred over the last century, many of them in the United States. The origins of testing, however, are neither recent nor American. Evidence suggests that the Chinese had a relatively sophisticated civil service testing program more than 4000 years ago (DuBois, 1970, 1972). Every third year in China, oral examinations were given to help determine work evaluations and promotion decisions.

By the Han Dynasty (206 B.C.E. to 220 C.E.), the use of **test batteries** (two or more tests used in conjunction) was quite common. These early tests related to such diverse topics as civil law, military affairs, agriculture, revenue, and geography. Tests had become quite well developed by the Ming Dynasty (1368–1644 C.E.). During this period, a national multistage testing program involved local and regional testing centers equipped with special testing booths. Those who did well on the tests at the local level went on to provincial capitals for more extensive essay examinations. After this second testing, those with the highest test scores went on to the nation's capital for a final round. Only those who passed this third set of tests were eligible for public office.

The Western world most likely learned about testing programs through the Chinese. Reports by British missionaries and diplomats encouraged the English

East India Company in 1832 to copy the Chinese system as a method of selecting employees for overseas duty. Because testing programs worked well for the company, the British government adopted a similar system of testing for its civil service in 1855. After the British endorsement of a civil service testing system, the French and German governments followed suit. In 1883, the U.S. government established the American Civil Service Commission, which developed and administered competitive examinations for certain government jobs. The impetus of the testing movement in the Western world grew rapidly at that time (Wiggins, 1973).

Charles Darwin and Individual Differences

Perhaps the most basic concept underlying psychological and educational testing pertains to individual differences. No two snowflakes are identical, no two fingerprints the same. Similarly, no two people are exactly alike in ability and typical behavior. As we have noted, tests are specifically designed to measure these individual differences in ability and personality among people.

Although human beings realized long ago that individuals differ, developing tools for measuring such differences was no easy matter. To develop a measuring device, we must understand what we want to measure. An important step toward understanding individual differences came with the publication of Charles Darwin's highly influential book, *The Origin of Species*, in 1859. According to Darwin's theory, higher forms of life evolved partially because of differences among individual forms of life within a species. Given that individual members of a species differ, some possess characteristics that are more adaptive or successful in a given environment than are those of other members. Darwin also believed that those with the best or most adaptive characteristics survive at the expense of those who are less fit and that the survivors pass their characteristics on to the next generation. Through this process, he argued, life has evolved to its currently complex and intelligent levels.

Sir Francis Galton, a relative of Darwin's, soon began applying Darwin's theories to the study of human beings (see Figure 1.5). Given the concepts of survival of the fittest and individual differences, Galton set out to show that some people possessed characteristics that made them more fit than others, a theory he articulated in his book *Hereditary Genius*, published in 1869. Galton (1883) subsequently began a series of experimental studies to document the validity of his position. He concentrated on demonstrating that individual differences exist in human sensory and motor functioning, such as reaction time, visual acuity, and physical strength. In doing so, Galton initiated a search for knowledge concerning human individual differences, which is now one of the most important domains of scientific psychology.

Galton's work was extended by the U.S. psychologist James McKeen Cattell, who coined the term *mental test* (Cattell, 1890). Cattell's doctoral dissertation was based on Galton's work on individual differences in reaction time. As such, Cattell perpetuated and stimulated the forces that ultimately led to the development of modern tests.

Experimental Psychology and Psychophysical Measurement

A second major foundation of testing can be found in experimental psychology and early attempts to unlock the mysteries of human consciousness through the scientific method. Before psychology was practiced as a science, mathematical models

FIGURE **1.5** Sir Francis Galton.

(From the National Library of Medicine.)

of the mind were developed, in particular those of J. E. Herbart. Herbart eventually used these models as the basis for educational theories that strongly influenced 19th-century educational practices. Following Herbart, E. H. Weber attempted to demonstrate the existence of a psychological threshold, the minimum stimulus necessary to activate a sensory system. Then, following Weber, G. T. Fechner devised the law that the strength of a sensation grows as the logarithm of the stimulus intensity.

Wilhelm Wundt, who set up a laboratory at the University of Leipzig in 1879, is credited with founding the science of psychology, following in the tradition of Weber and Fechner (Hearst, 1979). Wundt was succeeded by E. B. Titchner, whose student, G. Whipple, recruited L. L. Thurstone. Whipple provided the basis for immense changes in the field of testing by conducting a seminar at the Carnegie Institute in 1919 attended by Thurstone, E. Strong, and other early prominent U.S. psychologists. From this seminar came the Carnegie Interest Inventory and later the Strong Vocational Interest Blank. Later in this book we discuss in greater detail the work of these pioneers and the tests they helped to develop.

Thus, psychological testing developed from at least two lines of inquiry: one based on the work of Darwin, Galton, and Cattell on the measurement of individual differences, and the other (more theoretically relevant and probably stronger) based on the work of the German psychophysicists Herbart, Weber, Fechner, and Wundt. Experimental psychology developed from the latter. From this work also came the idea that testing, like an experiment, requires rigorous experimental control. Such control, as you will see, comes from administering tests under highly standardized conditions.

The efforts of these researchers, however necessary, did not by themselves lead to the creation of modern psychological tests. Such tests also arose in response to important needs such as classifying and identifying the mentally and emotionally handicapped. One of the earliest tests resembling current procedures, the Seguin Form Board Test (Seguin, 1866/1907), was developed in an effort to educate and

evaluate the mentally disabled. Similarly, Kraepelin (1912) devised a series of examinations for evaluating emotionally impaired people.

An important breakthrough in the creation of modern tests came at the turn of the 20th century. The French minister of public instruction appointed a commission to study ways of identifying intellectually subnormal individuals in order to provide them with appropriate educational experiences. One member of that commission was Alfred Binet. Working in conjunction with the French physician T. Simon, Binet developed the first major general intelligence test. Binet's early effort launched the first systematic attempt to evaluate individual differences in human intelligence (see Chapter 9).

The Evolution of Intelligence and Standardized Achievement Tests

The history and evolution of Binet's intelligence test are instructive. The first version of the test, known as the Binet-Simon Scale, was published in 1905. This instrument contained 30 items of increasing difficulty and was designed to identify intellectually subnormal individuals. Like all well-constructed tests, the Binet-Simon Scale of 1905 was augmented by a comparison or standardization sample. Binet's standardization sample consisted of 50 children who had been given the test under *standard conditions*—that is, with precisely the same instructions and format. In obtaining this standardization sample, the authors of the Binet test had norms with which they could compare the results from any new subject. Without such norms, the meaning of scores would have been difficult, if not impossible, to evaluate. However, by knowing such things as the average number of correct responses found in the standardization sample, one could at least state whether a new subject was below or above it.

It is easy to understand the importance of a standardization sample. However, the importance of obtaining a standardization sample that represents the population for which a test will be used has sometimes been ignored or overlooked by test users. For example, if a standardization sample consists of 50 white men from wealthy families, then one cannot easily or fairly evaluate the score of an African American girl from a poverty-stricken family. Nevertheless, comparisons of this kind are sometimes made. Clearly, it is not appropriate to compare an individual with a group that does not have the same characteristics as the individual.

Binet was aware of the importance of a standardization sample. Further development of the Binet test involved attempts to increase the size and representativeness of the standardization sample. A **representative sample** is one that comprises individuals similar to those for whom the test is to be used. When the test is used for the general population, a representative sample must reflect all segments of the population in proportion to their actual numbers.

By 1908, the Binet-Simon Scale had been substantially improved. It was revised to include nearly twice as many items as the 1905 scale. Even more significantly, the size of the standardization sample was increased to more than 200. The 1908 Binet-Simon Scale also determined a child's **mental age**, thereby introducing a historically significant concept. In simplified terms, you might think of mental age as a measurement of a child's performance on the test relative to other children

of that particular age group. If a child's test performance equals that of the average 8-year-old, for example, then his or her mental age is 8. In other words, in terms of the abilities measured by the test, this child can be viewed as having a similar level of ability as the average 8-year-old. The chronological age of the child may be 4 or 12, but in terms of test performance, the child functions at the same level as the average 8-year-old. The mental age concept was one of the most important contributions of the revised 1908 Binet-Simon Scale.

In 1911, the Binet-Simon Scale received a minor revision. By this time, the idea of intelligence testing had swept across the world. By 1916, L. M. Terman of Stanford University had revised the Binet test for use in the United States. Terman's revision, known as the Stanford-Binet Intelligence Scale (Terman, 1916), was the only American version of the Binet test that flourished. It also characterizes one of the most important trends in testing—the drive toward better tests.

Terman's 1916 revision of the Binet-Simon Scale contained many improvements. The standardization sample was increased to include 1000 people, original items were revised, and many new items were added. Terman's 1916 Stanford-Binet Intelligence Scale added respectability and momentum to the newly developing testing movement.

World War I

The testing movement grew enormously in the United States because of the demand for a quick, efficient way of evaluating the emotional and intellectual functioning of thousands of military recruits in World War I. The war created a demand for large-scale group testing because relatively few trained personnel could evaluate the huge influx of military recruits. However, the Binet test was an individual test.

Shortly after the United States became actively involved in World War I, the army requested the assistance of Robert Yerkes, who was then the president of the American Psychological Association (see Yerkes, 1921). Yerkes headed a committee of distinguished psychologists who soon developed two structured group tests of human abilities: the Army Alpha and the Army Beta. The Army Alpha required reading ability, whereas the Army Beta measured the intelligence of illiterate adults.

World War I fueled the widespread development of group tests. About this time, the scope of testing also broadened to include tests of achievement, aptitude, interest, and personality. Because achievement, aptitude, and intelligence tests overlapped considerably, the distinctions proved to be more illusory than real. Even so, the 1916 Stanford-Binet Intelligence Scale had appeared at a time of strong demand and high optimism for the potential of measuring human behavior through tests. World War I and the creation of group tests had then added momentum to the testing movement. Shortly after the appearance of the 1916 Stanford-Binet Intelligence Scale and the Army Alpha test, schools, colleges, and industry began using tests. It appeared to many that this new phenomenon, the psychological test, held the key to solving the problems emerging from the rapid growth of population and technology.

Achievement Tests

Among the most important developments following World War I was the development of standardized achievement tests. In contrast to essay tests, standardized achievement tests provide multiple-choice questions that are standardized on a

large sample to produce norms against which the results of new examinees can be compared.

Standardized achievement tests caught on quickly because of the relative ease of administration and scoring and the lack of subjectivity or favoritism that can occur in essay or other written tests. In school settings, standardized achievement tests allowed one to maintain identical testing conditions and scoring standards for a large number of children. Such tests also allowed a broader coverage of content and were less expensive and more efficient than essays. In 1923, the development of standardized achievement tests culminated in the publication of the Stanford Achievement Test by T. L. Kelley, G. M. Ruch, and L. M. Terman.

By the 1930s, it was widely held that the objectivity and reliability of these new standardized tests made them superior to essay tests. Their use proliferated widely. It is interesting, as we shall discuss later in the book, that teachers of today appear to have come full circle. Currently, many people favor written tests and work samples (portfolios) over standardized achievement tests as the best way to evaluate children (Boerum, 2000; Harris, 2002).

Rising to the Challenge

For every movement there is a countermovement, and the testing movement in the United States in the 1930s was no exception. Critics soon became vocal enough to dampen enthusiasm and to make even the most optimistic advocates of tests defensive. Researchers, who demanded nothing short of the highest standards, noted the limitations and weaknesses of existing tests. Not even the Stanford-Binet, a landmark in the testing field, was safe from criticism. Although tests were used between the two world wars and many new tests were developed, their accuracy and utility remained under heavy fire.

Near the end of the 1930s, developers began to reestablish the respectability of tests. New, improved tests reflected the knowledge and experience of the previous two decades. By 1937, the Stanford-Binet had been revised again. Among the many improvements was the inclusion of a standardization sample of more than 3000 individuals. A mere 2 years after the 1937 revision of the Stanford-Binet test, David Wechsler published the first version of the Wechsler intelligence scales (see Chapter 10), the Wechsler-Bellevue Intelligence Scale (W-B) (Wechsler, 1939). The Wechsler-Bellevue scale contained several interesting innovations in intelligence testing. Unlike the Stanford-Binet test, which produced only a single score (the so-called IQ, or intelligence quotient), Wechsler's test yielded several scores, permitting an analysis of an individual's pattern or combination of abilities.

Among the various scores produced by the Wechsler test was the performance IQ. Performance tests do not require a verbal response; one can use them to evaluate intelligence in people who have few verbal or language skills. The Stanford-Binet test had long been criticized because of its emphasis on language and verbal skills, making it inappropriate for many individuals, such as those who cannot speak or who cannot read. In addition, few people believed that language or verbal skills play an exclusive role in human intelligence. Wechsler's inclusion of a nonverbal scale thus helped overcome some of the practical and theoretical weaknesses of the Binet test. In 1986, the Binet test was drastically revised to include performance subtests. More recently, it was overhauled again in 2003, as we shall see in Chapter 9. (Other

important concepts in intelligence testing will be formally defined in Chapter 10, which covers the various forms of the Wechsler intelligence scales.)

Personality Tests: 1920–1940

Just before and after World War II, personality tests began to blossom. Whereas intelligence tests measured ability or potential, personality tests measured presumably stable characteristics or traits that theoretically underlie behavior. **Traits** are relatively enduring dispositions (tendencies to act, think, or feel in a certain manner in any given circumstance) that distinguish one individual from another. For example, we say that some people are optimistic and some pessimistic. Optimistic people tend to remain so regardless of whether or not things are going well. A pessimist, by contrast, tends to look at the negative side of things. Optimism and pessimism can thus be viewed as traits. One of the basic goals of traditional personality tests is to measure traits. As you will learn, however, the notion of traits has important limitations.

The earliest personality tests were structured paper-and-pencil group tests. These tests provided multiple-choice and true-false questions that could be administered to a large group. Because it provides a high degree of structure—that is, a definite stimulus and specific alternative responses that can be unequivocally scored—this sort of test is a type of structured personality test. The first structured personality test, the Woodworth Personal Data Sheet, was developed during World War I and was published in final form just after the war (see Figure 1.6).

As indicated earlier, the motivation underlying the development of the first personality test was the need to screen military recruits. History indicates that tests such as the Binet and the Woodworth were created by necessity to meet unique challenges. Like the early ability tests, however, the first structured personality test was simple by today's standards. Interpretation of the Woodworth test depended on the now-discredited assumption that the content of an item could be accepted at face value. If the person marked "False" for the statement "I wet the bed," then it was assumed that he or she did not "wet the bed." As logical as this assumption

	Yes	No
1. I wet the bed.	☐	☐
2. I drink a quart of whiskey each day.	☐	☐
3. I am afraid of closed spaces.	☐	☐
4. I believe I am being followed.	☐	☐
5. People are out to get me.	☐	☐
6. Sometimes I see or hear things that other people do not hear or see.	☐	☐

FIGURE **1.6** The Woodworth Personal Data Sheet represented an attempt to standardize the psychiatric interview. It contains questions such as those shown here.

seems, experience has shown that it is often false. In addition to being dishonest, the person responding to the question may not interpret the meaning of "wet the bed" the same way as the test administrator does. (Other problems with tests such as the Woodworth are discussed in Chapter 13.)

The introduction of the Woodworth test was enthusiastically followed by the creation of a variety of structured personality tests, all of which assumed that a subject's response could be taken at face value. However, researchers scrutinized, analyzed, and criticized the early structured personality tests, just as they had done with the ability tests. Indeed, the criticism of tests that relied on face value alone became so intense that structured personality tests were nearly driven out of existence. The development of new tests based on more modern concepts followed, revitalizing the use of structured personality tests. Thus, after an initial surge of interest and optimism during most of the 1920s, structured personality tests declined by the late 1930s and early 1940s. Following World War II, however, personality tests based on fewer or different assumptions were introduced, thereby rescuing the structured personality test.

During the brief but dramatic rise and fall of the first structured personality tests, interest in projective tests began to grow. In contrast to structured personality tests, which in general provide a relatively unambiguous test stimulus and specific alternative responses, projective personality tests provide an ambiguous stimulus and unclear response requirements. Furthermore, the scoring of projective tests is often subjective.

Unlike the early structured personality tests, interest in the projective Rorschach inkblot test grew slowly (see Figure 1.7). The Rorschach test was first published by Herman Rorschach of Switzerland in 1921. However, several years passed before the Rorschach came to the United States, where David Levy introduced it. The first Rorschach doctoral dissertation written in a U.S. university was not completed until 1932, when Sam Beck, Levy's student, decided to investigate the properties of the Rorschach test scientifically. Although initial interest in the Rorschach test was lukewarm at best, its popularity grew rapidly after Beck's work, despite suspicion, doubt, and criticism from the scientific community. Today, however, the Rorschach is under a dark cloud (see Chapter 14).

Adding to the momentum for the acceptance and use of projective tests was the development of the Thematic Apperception Test (TAT) by Henry Murray and Christina Morgan in 1935. Whereas the Rorschach test contained completely ambiguous inkblot stimuli, the TAT was more structured. Its stimuli consisted of ambiguous pictures depicting a variety of scenes and situations, such as a boy sitting in front of a table with a violin on it. Unlike the Rorschach test, which asked the subject to explain what the inkblot might be, the TAT required the subject to make up a story about the ambiguous scene. The TAT purported to measure human needs and thus to ascertain individual differences in motivation.

The Emergence of New Approaches to Personality Testing

The popularity of the two most important projective personality tests, the Rorschach and TAT, grew rapidly by the late 1930s and early 1940s, perhaps because of disillusionment with structured personality tests (Dahlstrom, 1969a). However, as we shall see in Chapter 14, projective tests, particularly the Rorschach, have not

FIGURE **1.7** Card 1 of the Rorschach inkblot test, a projective personality test. Such tests provide an ambiguous stimulus to which a subject is asked to make some response.

withstood a vigorous examination of their psychometric properties (Wood, Nezworski, Lilienfeld, & Garb, 2003).

In 1943, the Minnesota Multiphasic Personality Inventory (MMPI) began a new era for structured personality tests. The idea behind the MMPI—to use empirical methods to determine the meaning of a test response—helped revolutionize structured personality tests. The problem with early structured personality tests such as the Woodworth was that they made far too many assumptions that subsequent scientific investigations failed to substantiate. The authors of the MMPI, by contrast, argued that the meaning of a test response could be determined only by empirical research. The MMPI, along with its updated companion the MMPI-2 (Butcher, 1989, 1990), is currently the most widely used and referenced personality test. Its emphasis on the need for empirical data has stimulated the development of tens of thousands of studies.

Just about the time the MMPI appeared, personality tests based on the statistical procedure called *factor analysis* began to emerge. **Factor analysis** is a method of finding the minimum number of dimensions (characteristics, attributes), called *factors*, to account for a large number of variables. We may say a person is outgoing, is gregarious, seeks company, is talkative, and enjoys relating to others. However, these descriptions contain a certain amount of redundancy. A factor analysis can identify how much they overlap and whether they can all be accounted for or subsumed under a single dimension (or factor) such as extroversion.

In the early 1940s, J. R Guilford made the first serious attempt to use factor analytic techniques in the development of a structured personality test. By the end of that decade, R. B. Cattell had introduced the Sixteen Personality Factor Questionnaire (16PF); despite its declining popularity, it remains one of the most well-constructed structured personality tests and an important example of a test developed with the aid of factor analysis. Today, factor analysis is a tool used in the design

TABLE **1.2**
Summary of Personality Tests

Woodworth Personal Data Sheet: An early structured personality test that assumed that a test response can be taken at face value.
The Rorschach Inkblot Test: A highly controversial projective test that provided an ambiguous stimulus (an inkblot) and asked the subject what it might be.
The Thematic Apperception Test (TAT): A projective test that provided ambiguous pictures and asked subjects to make up a story.
The Minnesota Multiphasic Personality Inventory (MMPI): A structured personality test that made no assumptions about the meaning of a test response. Such meaning was to be determined by empirical research.
The California Psychological Inventory (CPI): A structured personality test developed according to the same principles as the MMPI.
The Sixteen Personality Factor Questionnaire (16PF): A structured personality test based on the statistical procedure of factor analysis.

or validation of just about all major tests. (Factor analytic personality tests will be discussed in Chapter 13.) See Table 1.2 for a brief overview of personality tests.

The Period of Rapid Changes in the Status of Testing

The 1940s saw not only the emergence of a whole new technology in psychological testing but also the growth of applied aspects of psychology. The role and significance of tests used in World War I were reaffirmed in World War II. By this time, the U.S. government had begun to encourage the continued development of applied psychological technology. As a result, considerable federal funding provided paid, supervised training for clinically oriented psychologists. By 1949, formal university training standards had been developed and accepted, and clinical psychology was born. Other applied branches of psychology—such as industrial, counseling, educational, and school psychology—soon began to blossom.

One of the major functions of the applied psychologist was providing psychological testing. The Shakow, Hilgard, Kelly, Sanford, and Shaffer (1947) report, which was the foundation of the formal training standards in clinical psychology, specified that psychological testing was a unique function of the clinical psychologist and recommended that testing methods be taught only to doctoral psychology students. A position paper of the American Psychological Association published 7 years later (APA, 1954) affirmed that the domain of the clinical psychologist included testing. It formally declared, however, that the psychologist would conduct psychotherapy only in "true" collaboration with physicians. Thus, psychologists could conduct testing independently, but not psychotherapy. Indeed, as long as psychologists assumed the role of testers, they played a complementary but often secondary role vis-à-vis medical practitioners. Though the medical profession could have hindered the emergence of clinical psychology, it did not, because as tester the psychologist aided the physician. Therefore, in the late 1940s and early 1950s, testing was the major function of the clinical psychologist (Shaffer, 1953).

For better or worse, depending on one's perspective, the government's efforts to stimulate the development of applied aspects of psychology, especially clinical psychology, were extremely successful. Hundreds of highly talented and creative young people were attracted to clinical and other applied areas of psychology. These individuals, who would use tests and other psychological techniques to solve practical human problems, were uniquely trained as practitioners of the principles, empirical foundations, and applications of the science of psychology.

Armed with powerful knowledge from scientific psychology, many of these early clinical practitioners must have felt frustrated by their relationship to physicians (see Saccuzzo & Kaplan, 1984). Unable to engage independently in the practice of psychotherapy, some psychologists felt like technicians serving the medical profession. The highly talented group of post-World War II psychologists quickly began to reject this secondary role. Further, because many psychologists associated tests with this secondary relationship, they rejected testing (Lewandowski & Saccuzzo, 1976). At the same time, the potentially intrusive nature of tests and fears of misuse began to create public suspicion, distrust, and contempt for tests. Attacks on testing came from within and without the profession. These attacks intensified and multiplied so fast that many psychologists jettisoned all ties to the traditional tests developed during the first half of the 20th century. Testing therefore underwent another sharp decline in status in the late 1950s that persisted into the 1970s (see Holt, 1967).

The Current Environment

During the 1980s, 1990s, and 2000s several major branches of applied psychology emerged and flourished: neuropsychology, health psychology, forensic psychology, and child psychology. Because each of these important areas of psychology makes extensive use of psychological tests, psychological testing again grew in status and use. Neuropsychologists use tests in hospitals and other clinical settings to assess brain injury. Health psychologists use tests and surveys in a variety of medical settings. Forensic psychologists use tests in the legal system to assess mental state as it relates to an insanity defense, competency to stand trial or to be executed, and emotional damages. Child psychologists use tests to assess childhood disorders. Tests are presently in use in developed countries throughout the world (Marsh, Hau, Artelt, Baumet, & Peschar, 2006; Black & William, 2007). As in the past, psychological testing remains one of the most important yet controversial topics in psychology.

As a student, no matter what your occupational or professional goals, you will find the material in this text invaluable. If you are among those who are interested in using psychological techniques in an applied setting, then this information will be particularly significant. From the roots of psychology to the present, psychological tests have remained among the most important instruments of the psychologist in general and of those who apply psychology in particular.

Testing is indeed one of the essential elements of psychology. Though not all psychologists use tests and some psychologists are opposed to them, all areas of psychology depend on knowledge gained in research studies that rely on measurements. The meaning and dependability of these measurements are essential to psychological research. To study any area of human behavior effectively, one must understand the basic principles of measurement.

In today's complex society, the relevance of the principles, applications, and issues of psychological testing extends far beyond the field of psychology. Even if you do not plan to become a psychologist, you will likely encounter psychological tests. Attorneys, physicians, social workers, business managers, educators, and many other professionals must frequently deal with reports based on such tests. Even as a parent, you are likely to encounter tests (taken by your children). To interpret such information adequately, you need the information presented in this book.

The more you know about psychological tests, the more confident you can be in your encounters with them. Given the attacks on tests and threats to prohibit or greatly limit their use, you have a responsibility to yourself and to society to know as much as you can about psychological tests. The future of testing may well depend on you and people like you. A thorough knowledge of testing will allow you to base your decisions on facts and to ensure that tests are used for the most beneficial and constructive purposes.

Tests have probably never been as important as they are today. For example, consider just one type of testing—academic aptitude. Every year more than 2.5 million students take tests that are designed to measure academic progress or suitability, and the testing process begins early in students' lives. Some presecondary schools require certain tests, and thousands of children take them each year. When these students become adolescents and want to get into college preparatory schools, tens of thousands will take a screening examination. Few students who want to go to a 4-year college can avoid taking a college entrance test. The SAT Reasoning Test alone is given to some 2 million high-school students each year. Another 100,000 high-school seniors take other tests in order to gain advanced placement in college.

These figures do not include the 75,000 people who take a special test for admission to business school or the 148,000 who take a Law School Admission Test—or tests for graduate school, medical school, dental school, the military, professional licenses, and others. In fact, the Educational Testing Service alone administers more than 11 million tests annually in 181 countries (Gonzalez, 2001). Nor do they include the millions of tests given around the world for research and evaluation purposes (Black & William, 2007; Marsh et al., 2006). As sources of information about human characteristics, the results of these tests affect critical life decisions.

SUMMARY

The history of psychological testing in the United States has been brief but intense. Although these sorts of tests have long been available, psychological testing is very much a product of modern society with its unprecedented technology and population growth and unique problems. Conversely, by helping to solve the challenges posed by modern developments, tests have played an important role in recent U.S. and world history. You should realize, however, that despite advances in the theory and technique of psychological testing, many unsolved technical problems and hotly debated social, political, and economic issues remain. Nevertheless, the prevalence of tests despite strong opposition indicates that, although they are far from perfect, psychological tests must fulfill some important need in the decision-making processes permeating all facets of society. Because decisions must be made, such tests will probably flourish until a better or more objective way of making decisions emerges.

Modern history shows that psychological tests have evolved in a complicated environment in which hostile and friendly forces have produced a balance characterized by innovation and a continuous quest for better methods. One interesting thing about tests is that people never seem to remain neutral about them. If you are not in favor of tests, then we ask that you maintain a flexible, open mind while studying them. Our goal is to give you enough information to assess psychological tests intelligently throughout your life.