The World of Psychological Testing

1

OBJECTIVES

1. List the major categories of tests, giving at least one example for each category.
2. Identify the major uses and users of tests.
3. Summarize major assumptions and fundamental questions involved in testing.
4. Outline significant features of the major periods in the history of testing.
5. Identify the six major forces influencing the development of testing.
6. Give a definition of a "test."

Introduction

This chapter provides an overview of the world of testing. Of course, everyone knows, at least roughly, what we mean by a "test" or "testing." Everyone has at least some familiarity with a variety of tests, for example, college admission tests, final examinations in courses, vocational interest inventories, and perhaps some personality measures. However, as we begin formal study of this world, it is important to develop both a more comprehensive and a more precise understanding of the field. "More comprehensive" so that we consider all types of tests and all relevant issues: We do not want to miss anything important. "More precise" so that we begin to acquire the technical expertise needed by professionals within the broader fields of psychology and allied disciplines: We will not be satisfied with just a passing acquaintance with these topics.

This is an ambitious agenda for one chapter. However, this opening chapter seeks only to provide an overview of these matters. The remaining chapters supply the details. There are a variety of ways to accomplish our goal of providing an overview and orientation to the field. No single way is best. We will use five perspectives or approaches to introduce the field, viewing it, as it were, from different angles or through different lenses. First, we outline the major categories of tests. Most of these categories correspond to chapters in the latter half of this book. In the process of describing these major categories, we mention examples of some of the more widely used tests. Second, we identify the major uses and users of tests. Who actually uses these tests and for what purposes? Third, we outline the primary issues that we worry about in testing. Notice that this outline—the list of principal worries—corresponds to the chapters in the first half of the book. Fourth, we trace the historical roots of contemporary testing. We mark off major periods in this history and identify some major forces that have shaped the field. Fifth, we examine some of the attempts to define *test*, *testing*, and some related terms. When we finish viewing the field through these five perspectives, we should have a good overview of the field of testing.

KEY POINTS SUMMARY 1.1 Five Ways to Introduce the Field

1. Categories of Tests
2. Uses and Users of Tests
3. Issues: Assumptions and Questions
4. Historical Periods and Forces
5. By Definition

Major Categories of Tests

We begin our exploration of the world of testing by identifying the major categories of tests. Any such classification is necessarily fuzzy around the edges. Categories often blend into one another rather than being sharply different. Nevertheless, an organizational scheme helps us to comprehend the breadth of the field. Key Points Summary 1.2 provides the classification scheme we use throughout the book. In fact, Chapters 8–15 follow this organization. This introductory chapter just touches on the major categories. Each category receives in-depth treatment later.

The first major division encompasses **cognitive ability tests**. In the world of psychological testing, the term *cognitive ability* includes a wide variety of cognitive functions, such as memory, spatial visualization, and creative thinking. Historically, the area has centered on intelligence, broadly defined. This category subdivides into individually administered cognitive ability tests, group-administered cognitive ability tests, and a variety of other ability tests, that is, other than intelligence tests. An example of an individually administered cognitive ability test is the *Wechsler Adult Intelligence Scale*,[1] abbreviated WAIS. Another classic example in this category is the *Stanford–Binet Intelligence Scale*. These tests are administered to individual examinees, one-on-one, by trained psychologists to provide an index of the overall mental ability of individuals. An example of a group-administered cognitive ability test is the *Otis–Lennon School Ability Test* (OLSAT). This test is administered to groups of students, usually in classroom settings, to gauge mental ability to succeed in typical school subjects. Another example of tests in this category is the SAT[2] used to predict success in college.

TRY IT!

To see how we cover a category in more depth later, flip to page 216. Quickly scan pages 216–222. You will see how subsequent chapters give details about tests mentioned in this opening chapter.

There are many other types of mental ability tests—nearly an infinite variety—including tests of memory, quantitative reasoning, creative thinking, vocabulary, and spatial ability. Sometimes these mental functions are included in the tests of general mental ability, but sometimes they are tested separately.

The next major category includes **achievement tests**. These tests attempt to assess a person's level of knowledge or skill in a particular domain. We cover here only professionally developed, standardized tests. We exclude the vast array of teacher-made tests used daily in the educational enterprise. Even excluding teacher-made tests, achievement tests are easily the most widely used of all types of tests. The first subdivision in this area includes achievement batteries used in

[1] In this opening chapter, we refer only to the first editions of tests. In subsequent chapters, we refer to the more recent editions and their corresponding initials, for example, WAIS-IV, MMPI-2, and so on.

[2] For many years, this test was titled the *Scholastic Aptitude Test*. The title was officially changed to the *Scholastic Assessment Test* in 1992 and later simply to the initials SAT. The old titles still appear in many publications.

KEY POINTS SUMMARY 1.2 Major Categories of Tests

- Cognitive Ability Tests
  - Individually Administered
  - Group Administered
  - Other Abilities
- Achievement Tests
  - Batteries
  - Single Subject
  - Certification, Licensing
  - Government-sponsored Programs
  - Individual Achievement Tests
  - Curriculum-based Measures
- Personality Tests
  - Objective Tests
  - Projective Techniques
  - Other Approaches
- Interests and Attitudes
  - Vocational Interests
  - Attitude Scales
- Neuropsychological Tests

elementary and secondary schools. Nearly everyone reading this book will have taken one or more of these achievement batteries. Examples include the *Stanford Achievement Test* and the *Iowa Tests*. All these batteries consist of a series of tests in such areas as reading, mathematics, language, science, and social studies. The second subdivision includes single-subject tests that cover only one area, such as psychology, French, or geometry. An example of such a test—one that many readers of this book have taken or will take—is the GRE[3]: Psychology Test.

The third subdivision includes the incredible variety of tests used for purposes of certification and licensing in such fields as nursing, teaching, physical therapy, airline piloting, and so on. None of the tests in this category is a household name. But they have important consequences for people in specific vocational fields.

Fourth, various government agencies sponsor certain achievement testing programs. Most prominent among these are statewide achievement testing programs in such basic subjects as reading, writing, and mathematics. In fact, such state assessment programs have assumed enormous importance in recent years as a result of new federal laws. In some states, high school graduation depends partly on performance on these tests. Other government-sponsored programs provide information about nationwide performance in a variety of areas. The best known of these efforts are the National Assessment of Educational Progress (NAEP) and the Trends in International Mathematics and Science Study (TIMSS), both of which are the subject of frequent reports in the media.

Fifth, there are individually administered achievement tests. The first four types of achievement tests are typically group administered. However, some achievement tests are individually administered in much the same way as individually administered mental ability tests. The individually administered achievement tests aid in the diagnosis of such conditions as learning disabilities. Finally, we have a relatively recent entry: curriculum-based measures.

The next major category includes the variety of tests designed to yield information about the human personality. The first subdivision includes what we call **objective personality tests**. In testing parlance, objective simply means the tests are objectively scored, based on items answered in a true–false or similar format. Examples of these objective personality tests are the *Minnesota Multiphasic Personality Inventory*, abbreviated MMPI, the *Beck Depression Inventory* (BDI), and the *Eating Disorder Inventory* (EDI). For both convenience and conceptual clarity, in subsequent chapters, we divide these objective tests into those designed to measure personality traits within the normal range and those designed as clinical instruments to measure pathological or disabling conditions.

[3] GRE used to be an acronym for Graduate Record Examination but now is a stand-alone designation, like SAT.

TRY IT!

Part of becoming a professional in this field involves learning the initials for tests. The initials are used routinely in psychological reports and journal articles, often without reference to the full name of the test. Become accustomed to this! Without referring to the text, see if you can give the full test title for each of these sets of initials:

EDI _____ WAIS _____ MMPI _____

The second major subdivision of personality tests includes **projective techniques**. With all these techniques, the examinee encounters a relatively simple but unstructured task. We hope that the examinee's responses will reveal something about his or her personality. The most famous of these techniques is the *Rorschach Inkblot Test*—sometimes just called the Rorschach, other times called the inkblot test. Other examples are human figure drawings, sentence completion techniques, and reactions to pictures. We include under personality measures a third category, simply labeled "other approaches," to cover the myriad of other ways psychologists have devised to satisfy our limitless fascination with the human personality.

The next major category of tests encompasses measures of interests and attitudes. The most prominent subdivision in this category includes **vocational interest measures**. These tests are widely used in high schools and colleges to help individuals explore jobs relevant to their interests. Examples of such tests are the *Strong Interest Inventory* (SII) and the *Self-Directed Search* (SDS). This category also includes numerous measures of attitudes toward topics, groups, and practices. For example, there are measures for attitude toward capital punishment, attitude toward the elderly, and so on.

Our final category includes **neuropsychological tests**. These are tests designed to yield information about the functioning of the central nervous system, especially the brain. From some perspectives, this should not be a separate category because many of the tests used for neuropsychological testing simply come from the other categories. Much neuropsychological testing employs ability tests and often uses personality tests, too. However, we use a separate category to capture tests used specifically to assess brain functions. Of special interest are tests of memory for verbal and figural material, psychomotor coordination, and abstract thinking.

TRY IT!

Here is a simple test used by neuropsychologists. It is called a Greek cross. Look at the figure for a moment. Then put it aside and try to draw it from memory. What behaviors and mental processes do you think are involved in this test?

Some Additional Ways to Categorize Tests

Thus far, we have categorized tests according to their predominant type of content. In fact, this is the most common and, from most perspectives, the most useful way to classify tests. However, there are a number of other ways to classify tests. We will list them briefly. See Key Points Summary 1.3.

KEY POINTS SUMMARY 1.3 Some Additional Ways to Categorize Tests

- Paper-and-Pencil versus Performance
- Speed versus Power
- Individual versus Group
- Maximum versus Typical Performance
- Norm-Referenced versus Criterion-Referenced

Paper-and-Pencil versus Performance Tests

In a **performance test**, the examinee completes some action such as assembling a product, delivering a speech, conducting an experiment, or leading a group. In a **paper-and-pencil test**, the examinee responds to a set of questions usually, as implied by the title, using paper and pencil. Many paper-and-pencil tests use multiple-choice, true–false, or similar item types. Traditional paper-and-pencil tests often appear now on a computer screen, with the answer marked by key stroke or mouse click.

Speed versus Power Tests

The essential purpose of a **speed (or speeded) test** is to see how fast the examinee performs. The task is usually quite simple. The person's score is how many items or tasks can be completed in a fixed time or how much time (e.g., in minutes or seconds) is required to complete the task. For example, how quickly can you cross out all the "e's" on this page? How quickly can you complete 50 simple arithmetic problems such as 42 + 19, 24 × 8, and so on? A **power test**, on the other hand, usually involves challenging material, administered with no time limit or a very generous limit. The essential point of the power test is to test the limits of a person's knowledge or ability (other than speed). The distinction is not necessarily all-or-none: pure speed versus pure power. Some power tests may have an element of speed. You can't take forever to complete the SAT. However, mental prowess and knowledge rather than speed are the primary determinants of performance on a power test. Some speed tests may have an element of power. You have to do some thinking and perhaps even have a plan to cross out all the "e's" on the page. However, crossing out "e's" is primarily a matter of speed, not rocket science.

Individual versus Group Tests

This distinction refers simply to the mode of test administration. An **individual test** can be administered to only one individual at a time. The classic examples are individually administered intelligence tests. An examiner presents each question or task to the individual and records the person's response. A **group test** can be administered to many individuals at the same time, that is, to a group. Of course, individuals receive their own scores from a group-administered test. In general, any group-administered test can be administered to one individual at a time, when circumstances warrant, but individually administered tests cannot be given to an entire group at once.

Maximum versus Typical Performance

Here is another useful distinction between types of tests. Some tests look for **maximum performance**. How well can examinees perform when at their best? This is usually the case with achievement and ability tests. On the other hand, we sometimes want to see a person's **typical performance**. This is usually the case with personality, interest, and attitude tests. For example, on a personality test we want to know how extroverted a person typically is, not how extroverted he can be if he is trying really hard to appear extroverted.

Norm-Referenced versus Criterion-Referenced

Many tests have norms based on performance of cases in a standardization program. For example, you may know that your score on the SAT or ACT is at the 84th percentile, meaning that you scored better than 84% of persons in the national norm group. This constitutes a **norm-referenced interpretation** of your performance on the test. In contrast, some test interpretations depend on reference to some clearly defined criterion rather than on reference to a set of norms. For example, an instructor may say: I want you to know all the key terms at the end of the chapter. On the instructor's test, you correctly define only 60% of the key terms, and this is considered inadequate—regardless of how well other people did on the test. This is **criterion-referenced interpretation**. Actually, it is the method of interpretation rather than the test itself that is either norm-referenced or criterion-referenced. We explore this distinction more fully in Chapter 3.

Uses and Users of Tests

A second way to introduce the world of testing is to identify the typical uses and users of tests. For the various categories of tests listed in the previous section, who actually uses these tests? What are the settings in which these tests are used? Consider these examples.

- John is a clinical psychologist in private practice. He sees a lot of clients suffering from anxiety and depression. Some cases may be mild, susceptible to short-term behavioral and cognitive-behavioral therapy. Others may be much more chronic where the presenting symptoms overlay a potentially schizophrenic condition. Early in his assessment of the clients, John routinely uses the MMPI and, for particularly perplexing cases, the Rorschach Inkblot Test.
- Kristen is a school psychologist. For students referred to her by teachers, she regularly reviews the school records containing scores from the *Otis–Lennon School Ability Test* and *Stanford Achievement Test*. In addition, she will administer the *Wechsler Intelligence Scale for Children* (WISC) and apply a behavior rating scale.
- Frank is a high school counselor. He supervises the school's annual administration of the *Strong Interest Inventory* (SII). Results of the test are distributed in homerooms. Frank cannot meet with every student about the SII results, but he prepares materials for homeroom teachers to help students interpret their reports.
- Annika is a developmental psychologist. She is interested in the stresses children undergo as they move from prepuberty to adolescence. In her longitudinal study, she uses a measure of self-concept (the *Piers-Harris Children's Self-Concept Scale*) to track changes in how participants feel about themselves. She also has intelligence test scores for the participants, taken from school records, simply to help describe the nature of her sample.
- Brooke is a neuropsychologist. In a product liability suit brought against an automobile manufacturer by an individual claiming to have sustained brain injury in an accident, Brooke, on behalf of the manufacturer, presents evidence garnered from a variety of tests that no brain injury occurred.
- Bill is assistant director of human resources for MicroHard, a company that hires nearly 100 new secretaries each year at its four different locations. Bill oversees the testing of 500 secretarial candidates per year. He tries to ensure that they have the skills, both technical and interpersonal, that will make them productive members of the "MicroHard team."
- Joe works for the State Department of Education. The legislature just adopted a bill requiring that all students pass tests in reading, mathematics, and writing to receive a high school diploma. Joe—lucky fellow—must organize preparation of these tests.

These are all examples of the typical uses and users of tests. Let us provide a more systematic catalog of the major uses and users of tests. As listed in Key Points Summary 1.4, we identify four major groups of users. There is considerable diversity within each group, but each group is relatively distinct in the way it uses tests. We also note that each group uses nearly all kinds of tests, as defined in the previous section, although certain types of tests predominate within each group.

KEY POINTS SUMMARY 1.4 Major Contexts for Use of Tests

1. Clinical
2. Educational
3. Personnel
4. Research

The *first* category includes the fields of clinical psychology, counseling, school psychology, and neuropsychology. We label all these applications as *clinical use*. In these professional applications, the psychologist is trying to help an individual who has (or may have) some type of problem. The problem may be severe (e.g., schizophrenia) or mild (e.g., choosing a college major). Testing helps to identify the nature and severity of the problem and, perhaps, provides some suggestions about how to deal with the problem. Testing may also help to measure progress in dealing with the problem.

A host of surveys have documented the extent of clinical test usage. An overview of the surveys shows that tests play a prominent role in the professional practice of psychology. We should add that for all these fields advanced training in the administration and interpretation of tests is required. Doctoral-level work in fields such as clinical, counseling, and school psychology typically entails several full courses in testing beyond the introductory work covered in this book.

A *second* major use of tests is in *educational settings*, apart from the clinical use that occupies the school psychologist or counselor. We refer here primarily to use of group-administered tests of ability and achievement. The actual users of the test information include teachers, educational administrators, parents, and the general public, especially as represented by such officials as legislators and school boards. Use of standardized testing in educational settings resolves into two major subdivisions. First, there are achievement tests used for determining levels of student learning. Limiting our counts to standardized achievement tests (i.e., excluding the vast array of teacher-made tests), tens of millions of these tests are administered annually. Achievement tests are also used to document competence for certification or licensing in a wide variety of professions.

The second primary use of tests in educational settings is to predict success in academic work. Prime examples in this category are tests for college and professional school admissions. For example, close to 2 million students take the SAT each year, while nearly 1 million students take the ACT. Approximately 300,000 GRE: General tests are administered annually, as are about 100,000 *Law School Admission Tests* (LSAT).

The *third* major category of test usage involves *personnel or employment* testing. Primary users in this category are businesses and the military. There are two essential tasks. The first task is to select individuals most qualified to fill a position. "Most qualified" usually means "most likely to be successful." For example, we may want to select from a pool of applicants the individuals who are most likely to be successful salespersons, managers, secretaries, or telemarketers. Tests may be useful in this selection process. The tests may include measures of general cognitive ability, specific job-related skills, and personality characteristics. Of course, nontest information will also be used. For example, letters of recommendation and records of previous employment are typical nontest sources of information.

The second task in the employment area has a different opening scenario. In the first case, we had a pool of applicants, and we selected from that pool. In the second case, we have a group of individuals who will be employed, and we need to assign them to different tasks to optimize the organization's overall efficiency. Tests may provide useful information about the optimum allocation of the human resources in this scenario.

The *fourth* major category of test usage is *research*. This is clearly the most diverse category. Tests are used in every conceivable area of research in psychology, education, and other social/behavioral sciences. For convenience, we can identify three subcategories of research usage. First, tests often serve as the dependent variable in a research study. More specifically, the test serves as the *operational definition of the dependent variable*. For example, in a study of the effects of caffeine on short-term memory, the *Wechsler Memory Scale* may be the operational definition of "memory." In a study of gender differences in self-concept, the *Piers-Harris Children's Self-Concept Scale* may provide the definition of self-concept. In a longitudinal study of the effects of an improved nutrition program on school performance, the *Stanford Achievement Test* may serve as the measure of performance. There are several major advantages to using an existing test as the operational definition of a dependent variable in such studies. First, the researcher need not worry about developing a new measure. Second, the existing test should have known properties such as normative and reliability information. Third and most important, use of an existing test helps replicability by other researchers.

The second major category of research usage is for purposes of *describing samples*. Important characteristics of the samples used in a research study should be delineated. The Method section of a research article often provides information about age and gender of participants. Some characteristics are described by test information—for example, means and standard deviations on an intelligence, achievement, or personality test. In a study of college students, it may be helpful to know the average SAT or ACT scores for the students. In a study of elderly patients in a state hospital, it may be helpful to know the patients' scores on the MMPI. Note that in these instances the test scores are not used as dependent variables but only to describe the research samples.

The third major category of research usage involves *research on the tests themselves*. As we will see in the next chapter, entire journals are devoted to this type of research. Furthermore, the development of new tests is itself a major research enterprise. Because tests play a prominent role in the social/behavioral sciences, continuous research on the tests is an important professional contribution.

Major Issues: Assumptions and Questions

A third important way to introduce the field of testing is to examine the fundamental issues, assumptions, and questions in the field. When psychologists think carefully about tests, regardless of the type of test, what issues do they worry about and what assumptions are they making? Describing these basic issues and assumptions helps us understand what the field is all about.

Basic Assumptions

Let us begin this way of exploring the field by identifying the assumptions we seem to make. There are probably four partly overlapping but reasonably distinct assumptions. First, we assume that human beings have recognizable *traits or characteristics*. Examples of traits are verbal ability, memory, extroversion, friendliness, quantitative reasoning ability, self-esteem, knowledge of Irish history, and depression. Furthermore, we assume that these traits or characteristics describe *potentially important* aspects of persons. More specifically, we assume that *differences* among

FIGURE 1.1
People differ. Tests can help gauge these differences.
Source: CartoonStock.com with permission.

individuals are potentially important. There are many ways in which people are the same. We all need oxygen. Without it, we quickly expire. We do not differ much from one another in that regard. Nearly everyone uses language to some extent, a distinctively human characteristic. However, we also differ from one another in certain ways. Some people are much taller than other people. Some people are more depressed than others. Some people are more intelligent. We assume that such differences among people in the traits we test are important rather than trivial. What do you think about the differences in Figure 1.1?

TRY IT!

We have just named a variety of human traits (verbal ability, depression, etc.). Try to name several more traits, some in the ability domain, some in the personality domain.

Ability traits: _____
Personality traits: _____

KEY POINTS SUMMARY 1.5 Four Crucial Assumptions

1. People differ in important traits.
2. We can quantify these traits.
3. The traits are reasonably stable.
4. Measures of the traits relate to actual behavior.

Second, we assume that we can *quantify* these traits. Quantification means arranging objects (in this case people) along a continuum. Think of a continuum as going from low to high or less to more. The continuum corresponds to the trait we are studying. At its most primitive level, quantification involves distinguishing among the objects on the continuum. The distinction may simply

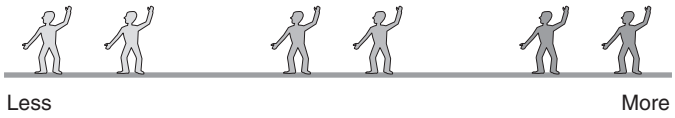


FIGURE 1.2
The continuum assumed for a trait.

be into two categories, labeled 0 and 1. At the next level of sophistication, we use the concept of "more or less" along the continuum, as shown in Figure 1.2. People are arrayed along the continuum for a trait. We examine these concepts of quantification in more detail in Chapter 3. For now, we simply note our assumption that such quantification of a trait is a fundamental notion in our work. It is this "quantification" assumption that gives rise to use of the term *measure* in the field of testing. In fact, in many contexts, "measure" is used as a synonym for "test." For example, the question "How did you measure the child's intelligence?" means the same as "How did you test the child's intelligence?"

Third, we assume that the traits have some degree of *stability or permanence*. They need not be perfectly stable, but they cannot fluctuate wildly from moment to moment. If the trait itself is insufficiently stable, then, no matter how refined our test, we will not be able to do much with it.

Fourth, we assume that the reasonably stable traits that we quantify with our tests have important *relationships to actual behavior* in real-life situations. From a theoretical perspective, this fourth assumption is the least important of our assumptions. That is, as theorists, we might be content to show that we can quantify a particular psychological trait regardless of whether it relates to anything else. However, from a practical perspective, this fourth assumption is crucial. As pragmatists, we might say that no matter how elegantly a test quantifies a psychological trait, if the test does not relate to anything else, we are not very interested in it.

Fundamental Questions

We now consider the fundamental questions about tests. In many ways, these questions are related to or are outgrowths of the assumptions previously listed. By way of anticipation, we note that these fundamental questions are precisely the topics covered in Chapters 3–7. In those chapters, we learn how psychologists try to answer these fundamental questions.

First, we ask about the **reliability** of the test. Reliability refers to the stability of test scores. For example, if I take the test today and again tomorrow, will I get approximately the same score? We examine this topic in considerable detail in Chapter 4. Notice that this question is not exactly the same as that treated in our third assumption. That assumption dealt with the stability of the trait itself. The question of reliability deals with the stability of our measurement of that trait.

Second, we ask about the **validity** of the test. Validity refers to what the test is actually measuring. If a test purports to measure intelligence, how do we know whether the test is, in fact, measuring intelligence? If a test purports to measure depression, how do we know that it is measuring depression? Included within the area of validity is the concept of **fairness**. Fairness, which is the flip side of bias, deals with the question of whether the test measures in an equitable manner across various groups, for example, across genders, ages, ethnic/racial groups, and different geographic areas. Such a question is, at root, a matter of the test's validity. We treat validity in detail in Chapter 5 and fairness in Chapter 7.

Third, we ask how to interpret the scores from a test. Olivia got a score of 13 right out of 20 items on the arithmetic test. Is that good or bad? Pete answered "True" to 45 of the 60 items on the depression scale. Does that mean he is depressed or positively euphoric? Interpretation of test scores usually depends on the use of **norms**. Norms are based on the test scores of large groups of individuals who have taken the test in the past. In Chapter 3, we describe the types of norms used with tests and how these norms are developed.

Questions related to reliability, validity, norms, and fairness are the most basic questions we ask about tests. Attempts to answer these questions form the core of test theory. These are the topics we worry about for all types of tests. However, we should add two additional types of questions to our catalog of fundamental questions. Knowing how a test was developed often enhances our understanding of reliability, validity, fairness, and norms. Hence, test development becomes another crucial topic. We consider it in Chapter 6. In addition, we need to consider a host of practical issues. How much does the test cost? How long does it take? Is it easily obtained? Is it available in languages other than English? All of these practical questions are important, although they are not part of test theory.

KEY POINTS SUMMARY 1.6 Fundamental Questions About Tests

- Reliability
- Validity
- Fairness
- Norms
- Test Development
- Practical Issues

The Differential Perspective

As a final note in the consideration of fundamental assumptions and questions, we call attention to what we will label the differential perspective. In many areas of the social and behavioral sciences, we attempt to formulate laws or generalizations that apply, more or less, to everyone. For example, what is the most effective Skinnerian schedule of reinforcement for learning a skill? What is the optimum level of stress for performing a certain task? Does psychoanalysis cure phobias? The formulation of such questions suggests that there is an answer that will generally hold true for people. In contrast, the **differential perspective** assumes that the answer may differ for different people. We are more interested in how people are different than in how they are the same. This differential perspective pervades the world of testing. Being aware of this perspective will help you think about issues in testing.

An emerging debate within the context of the differential perspective relates to how we think about differences: as *categories* or *dimensions* (see Widiger & Costa, 2012). A category describes a specific condition, such as having a broken ankle or a brain tumor. You either have it or you don't. The dimensional approach describes a continuum from less to more, low to high, or some similar set of descriptors. Think of height or speed in running a mile. Your height or speed is not something you do or don't have. All pretty clear examples. Now, what about depression? Is that a specific condition or simply the lower end of some continuum? What about learning disability? What about . . . ? Obviously, the list could go on. The categorical-versus-dimensional debate has important implications for how we think about the outcomes of psychological tests. The latest Diagnostic and Statistical Manual of Mental Disorders (DSM-5; American Psychiatric Association, 2013) includes several useful discussions of this difference in perspective.

The Historical Perspective

A fourth way to introduce the world of testing is by examining its historical origins. How did the field get where it is today? Knowing how we got here is often crucial for understanding today's issues. We provide this historical perspective in two ways. First, we outline major periods and events in the history of testing. Second, we sketch some of the major forces that have influenced

the development of testing. In constructing this history, we have drawn on a number of sources, many of which recount the same details but from somewhat different perspectives. For the earlier periods, see Boring (1950), DuBois (1970), Hilgard (1987), Misiak (1961), and Murphy (1949).

The history of testing can be conveniently divided into *seven major periods* (see Key Points Summary 1.7). Most of the periods have a dominant theme. The themes help to organize our understanding of the flow of events. We put chronological boundaries on the periods, with some rounding at the edges for pedagogical simplicity. We occasionally overstep our self-imposed boundaries to maintain continuity. In sketching the chronological development of the field, we avoid a mind-numbing recitation of dates, preferring to capture the spirit of different periods and transitions between these periods. We introduce a judicious selection of dates to highlight events that are particularly representative of a period. The reader will find it more useful to concentrate on the themes than on exact dates, although it is useful to commit a few dates to memory.

KEY POINTS SUMMARY 1.7 Major Periods in the History of Testing

1. Remote Background	Up to 1840	
2. Setting the Stage	1840–1880	40 years
3. The Roots	1880–1915	35 years
4. Flowering	1915–1940	25 years
5. Consolidation	1940–1965	25 years
6. Just Yesterday	1965–2000	35 years
7. And Now	2000–Present	

Remote Background: Up to 1840

The first period is rather artificial. It is so long as nearly to defy any meaningful summary. But we do need to start somewhere. Let us identify three noteworthy items in this broad expanse of time. First, we observe that the remote roots of psychology, as well as most fields, are in philosophy. Among the classical thinkers of ancient, medieval, and more modern times, there was a distinct lack of interest in the topic of individual differences or any notion of measuring traits. If we use the modern method of "citation frequency" to define authors' influence of the past 2,500 years, no doubt Aristotle, Plato, and Aquinas would emerge as the top three (beyond the sacred scriptures). Examination of the writings of these three giants, as well as of their colleagues, reveals a dominant interest in defining what is common to human beings, what is generally true, rather than what is different about them. Consider, for example, Aristotle's *Peri Psyche* (also known by its Latin name *De Anima*, translated into English as *On the Soul*). Written about 350 B.C., the work is often cited as the first textbook on psychology, indeed essentially giving the name to the field. In the opening book of his treatise, Aristotle (1935) says, "We seek to examine and investigate first the nature and essence of the soul, and then its [essential] attributes" (p. 9). This is not the stuff of the differential perspective.

Plato, the other great luminary of the ancient world, whose influence also continues unabated, similarly concentrated on the general and, even more so than Aristotle, on the abstract. The most influential writer of the medieval period was Thomas Aquinas. With respect to psychological matters, he recapitulated much of Aristotle's work. Indeed, he saw his principal task as that of reconciling Christian theology with the Aristotelian synthesis. Hence, Aquinas adopts Aristotle's concepts of human abilities and manifests the same disinterest in human differences, preferring to concentrate on general characteristics of human nature. Of course, these philosophers were not

fools. They were, in fact, keen observers of the human condition. They have occasional—often fascinating—comments on matters of individual differences, but such comments are strictly sidelights, not a focus of attention.

Following the medieval period, the Renaissance witnessed a true awakening to the individual. But this interest was reflected primarily in artistic productions, the glorious profusion of paintings, sculptures, and buildings that still leaves us breathless. Dominant thinkers of the late Renaissance and early modern period continued to concern themselves with how the human mind works. For example, Descartes, Locke, Hume, and Kant framed questions—and gave answers—that formed part of the remote background for psychology's roots. Emphasis continued to be placed on what was common.

Regarding the mode of examinations in our period of the remote past, DuBois (1970) observes that written examinations were not common in the Western educational tradition. The more usual practice in the schools throughout ancient, medieval, and, indeed, up until the mid-1800s was oral examination. The vestiges of this practice remain today in the oral defense of an honors, master's, or doctoral thesis, popularly known as "taking your orals," as if it were some kind of distasteful pill (actually, it is much worse than any pill). DuBois notes that written examinations emerged in the remarkable string of Jesuit schools in the late 1500s, the predecessors of today's network of Jesuit secondary schools, colleges, and universities throughout the world. The Jesuit *Ratio Studiorum*, a kind of early curriculum guide, laid down strict rules (standardization!) for conducting written examinations.

Finally, some textbooks report the equivalent of civil service examinations being used routinely in China as early as 2000 B.C. However, Bowman (1989) argues convincingly that these reports are based on inadequate (we might say near apocryphal) historical sources and that the earliest such testing probably occurred around 200 B.C. Nevertheless, whether 200 or 2200 B.C., this is an interesting historical development. The system continued until the early twentieth century and may have had some influence on civil service testing in Western countries.

Setting the Stage: 1840–1880

Events in the years 1840–1880 set the stage for the stars who were to be the main characters in the drama that unfolded next. This stage-setting was a largely disconnected set of events. However, in retrospect, we can see four strands coming together.

First, throughout this period, both scientific interest and public consciousness of mental illness increased enormously. From the early prodding of Philippe Pinel in France, Samuel Tuke in England, and Benjamin Rush in the United States, a host of efforts to improve the diagnosis and treatment of the mentally ill arose. Dorothea Dix (Figure 1.3) epitomized the humanitarian side of the effort. Beginning around 1840, she conducted a virtually worldwide crusade, resulting in improvements in prison and hospital conditions. On the scientific side, methods for the diagnosis of mental illness, including intellectual disability began to emerge. For example, simple methods of assessing mental ability, such as the Seguin form board (see Figure 1.4), appeared. These early measures had no norms, no reliability data. But they foreshadowed at least elements of measures that would develop later.

A second significant development of this period was the adoption of formal written examinations by the Boston School Committee—essentially the city's school board—under the direction of Horace Mann, probably the most influential educator of the day. Mann advocated, not only in Boston but nationwide, for drastic improvement in the way schools evaluated their students.

Third, the age of Darwin dawned. The world-shattering *On the Origin of Species by Means of Natural Selection* appeared in 1859. Perhaps more important for the incipient field of psychology were Darwin's subsequent books: *The Descent of Man, and Selection in Relation to Sex* in 1871 and *The Expression of the Emotions in Man and Animals* in 1872. Of course, all these turned the

United States Library of Congress's Prints and Photographs division (public domain)

FIGURE 1.3
Dorothea Dix, crusader for improvement in hospital conditions.

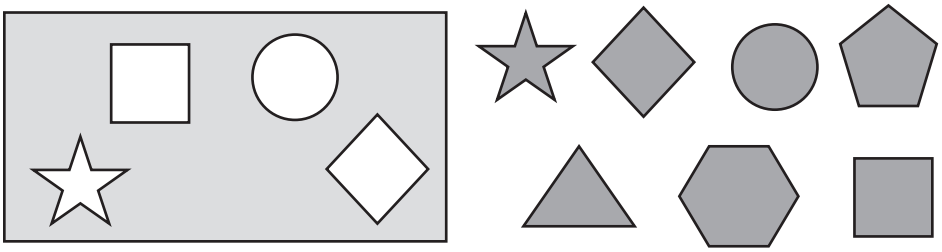


FIGURE 1.4
A Seguin-like form board.

world on its ear. But why were these works important for psychology? Because they got people thinking about differences: first differences between species, then differences between individuals. Most specifically, the works got Francis Galton thinking about individual differences. More on Galton in a moment.

Fourth, experimental psychology emerged. The traditional birthdate for the field is 1879, the year Wilhelm Wundt opened his laboratory at the University in Leipzig, now a city of one-half million people located 75 miles south of Berlin, Germany. Experimental psychology was an outgrowth of physiology. The connecting link was psychophysics. Early experimental psychology was essentially synonymous with psychophysics. Its ultimate contribution to the world of testing, for good or ill, was twofold. First, like any good laboratory science, it concentrated on standardization of conditions and precision of measurement. Second, it concentrated on elemental processes: sensation, thresholds, perception, simple motor reactions, and so on. Wundt's laboratory was the training ground of choice for many early psychologists. Hence, Wundt's interests and methods exercised great influence on the fledgling field of psychology.

KEY POINTS SUMMARY 1.8 Strands in Setting the Stage

- Increased interest in mental illness
- Adoption of written examinations
- Dawning of the age of Darwin
- Emergence of experimental psychology

Thus, we come to about 1880. Experimental psychology is a new science. The world is abuzz about evolution. There is widespread interest in the mentally ill. And some pioneers are trying to bring education into the scientific fold.

The Roots: 1880–1915

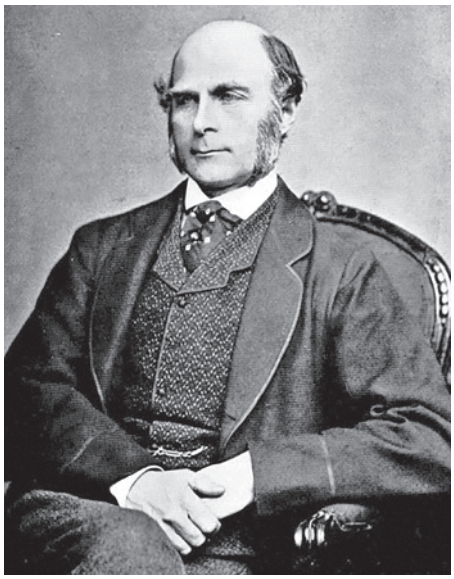
The roots of testing, as the enterprise exists today, were set in the 35-year period 1880–1915. The earliest measures that had lasting influence emerged in this period. Many of the basic issues and methodologies surfaced in more or less explicit form. At the beginning of the period, there were few—very few—examples one could point to and say: That's a test. By the end of the period, there was a panoply of instruments. Some of these, minus a few archaic words, are indistinguishable from today's tests. At the beginning of the period, the correlation coefficient and the concept of reliability had not been invented. By the end of the period, these methodological cornerstones of testing had not only been invented but had been elaborated and incorporated into an emergent theory of mental tests.

Let us highlight the key events and personalities of this exciting period. The highlights center around four key individuals. In addition, we will mention one other individual and then a loose confederation of other contributors.

The first key figure was **Francis Galton** (Figure 1.5). Many people consider him the founder of psychological testing. An independently wealthy British gentleman, he never had a real job, not even as a university professor. He dabbled. But when he did so, he did it in grand style, with astonishing creativity and versatility.

Fittingly, as a second cousin of Charles Darwin, Galton was the primary pipeline for transmitting evolution into the emerging field of psychology. Galton's interest lay in heredity, especially the inheritance of high levels of ability. He called it "genius" and studied it in a wide array of fields, including music, military and political leadership, and literature.

In trying to examine the relationships among the many variables he studied, Galton invented the bivariate distribution chart. As a follow-up, he induced Karl Pearson, a contemporary British mathematician, to invent the correlation coefficient. Galton had the time, resources, and personality to accomplish much. In addition, he was a proselytizer. He spread the word about methods of mental measurement. Despite the fact that he held no prestigious position, it seemed that by 1910 everyone knew about Galton's work.



Wikimedia Commons (public domain)

FIGURE 1.5
Francis Galton: dabbler extraordinaire and pipeline for evolutionary theory into psychology.

The key American contributor to the development of testing was **James McKeen Cattell**. After a brief stint at the University of Pennsylvania, he spent most of his professional career at Columbia University in New York City. Cattell's preparation was ideal for merging two methodological streams. He completed graduate work first with Wundt at Leipzig, honing his skills in rigorous laboratory studies of the psychophysical tradition. He then studied with Galton, apparently absorbing Galton's fascination with collecting data about individual differences in human traits. Consistent with the prevailing notion at the time, Cattell believed that the key to mental functioning was elemental processes. He created a battery of 50 tests, 10 of which were considered mainstays (see Table 1.1), covering such areas as sensory acuity, reaction time, visual bisection of a line, and judgments of short intervals of time. He administered these to groups of college students in hopes of predicting academic success—the conceptual grandparent of today's SATs and ACTs. He persuaded other psychologists to undertake similar projects. Cattell's tests were a colossal flop as predictors. Nevertheless, the work was highly influential. In a famous article of 1890, he coined the term *mental test* (Cattell, 1890), a term used to characterize the field for the next 50 years. Appropriately, a commentary by Galton followed the article.

TRY IT!

Which of the tests listed in Table 1.1 do you think might be the best predictors of academic success?

The third influential figure of this period was the Frenchman **Alfred Binet** (pronounced Bă-nay′). In his section on Mental Tests, Boring (1950) summarizes the matter succinctly: "The 1880s were Galton's decade in this field, the 1890s Cattell's, and the 1900s Binet's" (p. 573). Binet is truly the father of intelligence testing. Oddly, his original training was in the law, although he subsequently completed advanced training in medicine and natural sciences. Throughout most of his career, Binet concentrated on investigation of mental functions. In contrast to Galton, Binet aimed at more holistic mental activities: using words, finding connections, getting the meaning, and so on. The Parisian schools at that time wanted to identify students more likely to profit from special schools than from regular classroom instruction. A committee, including Binet and Theodore Simon, set out to devise a method to identify these students. The result was the Binet–Simon Scale,[4] first published in 1905. It appeared in revised form, including the first use of "mental ages," in 1908 and again in 1911. In Chapter 8, we examine the grandchild of Binet's scale, the modern-day *Stanford–Binet Intelligence Scale*.

Table 1.1 An Abbreviated List of Cattell's Ten Key Tests

1. Dynamometer Pressure [grip-strength]
2. Rate of Movement
3. Sensation-areas
4. Pressure causing Pain
5. Least noticeable difference in Weight
6. Reaction-time for Sound
7. Time for naming Colours
8. Bi-section of a 50 cm. Line
9. Judgment of 10 seconds Time
10. Number of letters remembered on once Hearing

[4]We follow the modern practice of referring to this as the Binet-Simon Scale. In his own work, Binet did not use an official name for it. He simply referred to it as the "test" or the "scale."

KEY POINTS SUMMARY 1.9 Important Persons in Establishing the Roots

- Francis Galton
- James McKeen Cattell
- Alfred Binet
- Charles Spearman
- Developers of "New Tests"

Fourth, there was the work of **Charles Spearman**, another Englishman. His contributions were of a distinctly different character from those of the first three. Spearman did not invent any new types of tests or test items. He did not undertake any novel data collection projects. Spearman was a grand theorist and a number cruncher. In 1904, he published a paper announcing the "two-factor" theory of intelligence. Oddly, it appeared in the *American Journal of Psychology* rather than in a British outlet. Spearman undergirded his theory with the method of tetrad differences, the earliest form of the statistical technique eventually known as factor analysis. The important point here is that this was the first attempt to provide an empirically based theory of human intelligence, and was an outcropping of the new methods of mental measurement. Previous theories were essentially philosophical. Here, based on test results (mostly from Galton-type tests), was a new kind of theory, accompanied by an entirely new mathematical methodology. This, indeed, was the stuff of a new science.

The final element in establishing the roots of testing is not so neatly identified. It was not just one person. Rather, it was a whole gaggle of persons, all pursuing the same goal in much the same way. This was the group of people feverishly building the earliest versions of educational achievement tests. They responded to much the same impulse as Horace Mann's: the need to bring education into the scientific world. If education was to be conducted in a scientific manner, then it needed precise, reliable measures. This was a different interest from Cattell's interest in prediction. These investigators wanted measures of the outcomes of education.

With near evangelistic zeal, a bevy of test authors created what they called "new-type" achievement tests. The principal concern was with the unreliability of essay and oral exams. The new-type tests were as objective as possible. In practice, this meant multiple-choice, true-false, and fill-in-the-blank test items. These items could be objectively scored, and they were more reliable than the "old-type" tests. The literature of the day is replete with references to the inadequacies of old-type tests. In today's parlance, the problem was interscorer reliability. Many people suppose that multiple-choice and similar types of items were invented to accommodate mass processing of tests by computers. Nothing could be more absurd. Computers did not even exist when the new-type tests emerged. Nor did answer-sheet scanners exist. Authors laboring in the achievement test field at this time were not concerned with scoring efficiency. Scoring reliability was their passion.

The Flowering: 1915–1940

From its humble and rather disjointed roots of 1880–1915, testing entered a period of spectacular growth. At the beginning of this period, there were tests but few were standardized in the manner we think of such instruments today. By the end of the period, in a mere 35 years, thousands of such tests were available. The first editions of the great majority of tests that are widely used today sprang up during this period. This happened in every sphere of testing: mental ability, achievement, personality, interests, and so on. The extent of activity was simply breathtaking. Let us examine some of these developments.

The clear demarcation point between the period of roots and the period of flowering comes when Binet's scales crossed the Atlantic Ocean from France to the United States. We might date this anywhere between 1910 and 1916. Whatever the specific date chosen, the important point

is the transatlantic crossing. Binet's work received almost immediate attention in the United States. Some of the new American versions were mainly translations, perhaps the earliest being Goddard's in 1910 (DuBois, 1970; Murphy, 1949). There were other translations and adaptations, too. However, the definitive event was publication of the Stanford Revision of the Binet Scales in 1916, popularly referred to as the Stanford–Binet. Engineered by Lewis Terman of Stanford University, the Stanford Revision involved new items (nearly doubling the original number), new tryout research, an ambitious national norming program, and the first widespread use of the ratio IQ—altogether a blockbuster. A researcher looking back on it today would scoff. At the time, it was like the first jet-powered aircraft, the first man on the moon, the first smart phone. Within a relatively short time, the Stanford–Binet became the benchmark definition of human intelligence, a mainstay of clinical practice, and perhaps the most distinctive symbol of psychology's contribution to the modern world. Thus began the period of flowering.

One of the most profoundly influential events in the history of testing was development of the first widely used group-administered intelligence test. This occurred in the context of psychologists' efforts to aid in processing the tidal wave of military recruits as the United States entered World War I in 1917. Arthur Otis, as part of his doctoral work under Lewis Terman (of Stanford–Binet fame), undertook creation of a group-administered form of the Stanford–Binet. Otis's work eventuated in the Army Alpha and Beta, verbal and nonverbal versions, respectively, administered to nearly 2 million military personnel. In 1918, the tests were made available for general use as the *Otis Group Intelligence Scale*. We will examine the lineal descendant of this test, the *Otis–Lennon School Ability Test*, in Chapter 9.

The Stanford–Binet and Otis tests established the use of a single score to represent intelligence. The major challenge to this practice arose in the work of L.L. Thurstone (1938), who maintained that there were seven (more or less) different dimensions of human intelligence. Thurstone's work spawned a host of multiscore intelligence tests in this period.

A flurry of publications in the relatively brief span of ten years, 1921–1930, established the preference for the "new-type" achievement test (McCall, 1922; Odell, 1928; Ruch, 1924, 1929; Ruch & Rice, 1930; Ruch & Stoddard, 1927; Toops, 1921; Wood, 1923). Although the previous era witnessed the earliest development of a host of "new-type" achievement tests, none gained widespread usage. The first truly national standardized achievement test was the *Stanford Achievement Test*, which appeared in 1923. Interestingly, one of its coauthors was Lewis Terman, principal architect of the Stanford revision of the Binet. The 1930s also witnessed the origin of several other well-recognized achievement batteries, as well as a host of single-subject tests in every conceivable area.

Personality testing, both objective and projective, also flourished during this period. The prototype of today's objective personality inventories, the *Woodworth Personal Data Sheet*, was devised to help process military recruits for World War I. It was essentially a paper-and-pencil interview, 116 items in all, each answered "Yes" or "No," to help detect individuals needing more thorough psychological examination. A host of similar instruments sprouted after World War I. Illustrating the profusion of new publications during this period are the following.

- Rorschach's inkblots appeared in 1921. By 1940, there were several different scoring systems for the inkblots.
- Strong and Kuder launched their pioneering work on vocational interest inventories (see Donnay, 1997; Zytowski, 1992); we describe the current versions of those tests in Chapter 15.
- The MMPI (see Chapter 13), though not appearing until just beyond the boundary of the period, was conceptualized during this time.
- Thurstone and Likert first attempted systematic measurement of attitudes; we describe their methods, still used today, in Chapter 15.

Earlier we recounted the benchmark event at the onset of this period of flowering: publication of the *Stanford–Binet Intelligence Scale* in 1916. Perhaps fittingly, the end of this period witnessed the arrival of the first major revision of this test in 1937. Almost coincidentally, in 1939, the *Wechsler–Bellevue Intelligence Scale* appeared. David Wechsler, a clinical psychologist working at New York City's Bellevue Hospital, was unhappy about using the Stanford–Binet—a test designed for children—with his adult clients. He created his test to be more suitable for adults.

First editions of three publications served like triple exclamation points near the conclusion of this remarkably fecund period in the history of testing. First was the publication of the highly theoretical journal *Psychometrika* in 1936; second was publication of the more pragmatically oriented *Educational and Psychological Measurement* in 1941; third was the first edition of Oscar K. Buros's *Mental Measurements Yearbook* in 1938, a work whose current edition we describe in detail in Chapter 2.

Consolidation: 1940–1965

Following the burst of activity on a multitude of fronts from 1915 to 1940, testing entered a period that might best be characterized as consolidation or maturity. This period lasted roughly from 1940 to 1965, another span of 25 years. Activity did not diminish—indeed, it continued to flourish. New, revised editions of many of the tests first developed in the previous period now appeared. Entirely new tests were also developed. Testing expanded in clinical practice, in the schools, in business, and in the military. But testing was no longer a new kid on the block. It was accepted professional practice. It was assumed that tests would play a prominent role in a variety of venues. A number of events signaled this newfound maturity.

Early in this period, of course, World War II (1939–1945) preoccupied everyone's attention. Tests, rather than being created anew as in World War I, were used widely and routinely for processing military personnel. Prominent psychologists, now trained in the testing methods developed in the previous period, guided these applications. In addition, clinical psychologists plied their trade in the treatment of war-related psychological casualties, in part using the tests now available.

The appearance of books or other written documents often defines a historical period. Thus, the Declaration of Independence signified the emergence of a new nation—although many other events could easily be taken as more important developments. Perhaps the best evidence of the consolidation of the field of testing during the period 1940–1964 was the appearance of a number of books summarizing the status of testing. These became classics in the field precisely because they could provide summaries of mature thinking about the major issues in the field. Twenty years earlier, say in 1930, it would not have been possible to write these books because thinking had not matured regarding the major issues.

Embedded within this period of consolidation was a particularly remarkable 6-year span, 1949–1954, when a half-dozen soon-to-be classics appeared. Among these works were the earliest versions (in 1954 and 1955) of what would become the *Standards for Educational and Psychological Tests*, a sort of bible of the best thinking on technical matters related to tests. We cite excerpts from these *Standards* throughout later chapters.

In 1950, Harold Gulliksen's *Theory of Mental Tests*, the then-definitive work on psychometric theory, appeared. At about the same time, the first editions of two seminal textbooks on testing emerged: Lee Cronbach's *Essentials of Psychological Testing* in 1949 and Anne Anastasi's *Psychological Testing* in 1954. (See Figure 1.6.) Both books subsequently appeared in a number of revised editions, but these first editions helped to define a mature field of study. Thus, testing entered the 1960s with a wide array of instruments, established patterns of usage, a well-defined theoretical base, and benchmark publications summarizing all of this.


Chuck Painter/Stanford News Service


Photo courtesy of Jonathan Galente

FIGURE 1.6
Authors of early textbooks on testing: Lee Cronbach and Anne Anastasi.

Just Yesterday: 1965–2000

Someone reading this book 50 years hence—if anyone does—will, no doubt, chuckle over a period labeled "just yesterday." Nevertheless, at this writing, the 35-year period 1965–2000 is "just yesterday." Four significant topics appear to characterize the period 1965–2000.

First, test theory changed dramatically. The period of consolidation essentially summarized what we now call **classical test theory**. The mid-1960s saw the emergence of **item response theory** or "modern test theory," a new set of methods for examining a whole range of issues related to the reliability, scaling, and construction of tests. The onset of the new theoretical approach is perhaps best signaled by the publication of *Statistical Theories of Mental Test Scores* by Frederic Lord and Melvin Novick (1968), that is, just at the beginning of the current period. This book, billed as the successor to Gulliksen's *Theory of Mental Tests*, ushered in a new era in test theory. Throughout the 1970s and continuing to the present, journals and textbooks devoted to testing exploded with applications of item response theory. We will refer to these developments in more detail in subsequent chapters.

Second, the mid-1960s witnessed the origins of both legislative and judicial activism regarding tests, emanating principally but not exclusively from the federal government. Heretofore, testing was not legislated—either for or against. Now, some types of tests were being required by law, while other types of tests, or certain uses of tests, were being prohibited. To say the least, this period of legislative and judicial activism presented a unique set of challenges. We examine specific cases arising within this milieu in Chapter 16.

Third, testing became the subject of widespread public criticism during this period. The criticisms were directed primarily at standardized tests of ability and achievement. Tests of interests, attitudes, and personality were mostly unscathed. During the 50 years before this period, testing was largely viewed as a new, valuable scientific tool. To be sure, there were debates, but they were mostly squabbles within the family, confined to the academic journals in psychology and education.

KEY POINTS SUMMARY 1.10 Significant Features: Just Yesterday

- Emergence of Item Response Theory
- Legislative and Judicial Activism
- Public Criticism of Testing
- Influence of Computers

Beginning in the mid-1960s, however, criticisms originated outside the field. They also intensified within the field, going much beyond the squabble stage. If you like controversy, testing is a good field to be in today! Subsequent chapters will help us sort through all of these criticisms and examine how to analyze them.

Fourth, computers have pervasively influenced contemporary testing. It may come as a surprise to today's reader, but this influence is very recent. The roots of testing, its flowering, and its period of consolidation all preceded the computer age. However, in the last 30 to 40 years much about the current practice of testing has changed as a result of computers. We reserve the telling of this tale to our discussion of major forces influencing testing in the next section.

And Now: 2000–Present

To say that we are writing a history of the present is an oxymoron. It is also a very dangerous one. Mistaking a temporary blip for a major trend, or having a blindspot for the emergence of a truly significant trend, can make one look foolish. Nevertheless, we conclude this sketch of the history of testing by identifying what appear to be notable developments in the current scene. We identify five such developments, prefacing the discussion by noting that the four trends from the previous period are still very much alive, and that what we label as current developments are outcroppings from the previous period.

First, there is the *explosive increase* in the number and diversity of tests. Every day brings the announcement of new instruments or revisions in existing editions. This growth phenomenon seems to be affecting all areas of psychological and educational testing. A notable subdivision of the activity occurs in statewide assessment programs, driven primarily by recent federal laws. As a result, in effect every state is now a test developer/publisher. But we are also witnessing a profusion of new tests of personality, various disorders, and mental abilities. Even cataloging all the new entries has become a nearly impossible task, and evaluating them in a timely fashion presents a daunting challenge.

This meteoric growth emphasizes the need for proficiency in using sources of information about tests—precisely the topic of our next chapter. It also underscores the need to be competent in evaluating the plethora of new tests, competence we aspire to develop in Chapters 3–7.

Second is the pervasive influence of *managed care*. Of course, managed care did not start in 2000, but it was unheard of until the latter part of the previous period. It is now one of the most influential forces in clinical practice. Managed care exerts pressure on testing in several ways. It calls for more focused testing: Don't use a two-hour, omnibus battery if you can get by with a 15-minute, more focused test. Managed care also demands careful links between diagnosis and treatment, on the one hand, and between treatment and outcomes, on the other hand. Therefore, the test results should point to treatment and treatment outcomes should be documented. In practice, that means repeated use of a test to show improvement: defined as change in test score.

Third, an outgrowth of the scientist-practitioner model described later, is the emergence of **evidence-based practice** (EBP)—the notion that whatever the psychologist does in practice should be based on sound evidence. As noted by Norcross et al. (2017), "Since the early 1990s, we have witnessed impressive growth in the number of articles invoking EBPs.... Truly, EBP has become an international juggernaut" (p. 2). Psychological tests play a crucial role in EBP. Much of the "evidence" in EBP comes from tests—such as those covered later in this book. Furthermore, understanding that evidence requires precisely the kind of knowledge developed in our next six chapters.

The fourth and fifth areas of development both relate to computers but in very different ways. The next section (Major Forces) traces the long-term influence of computers on testing. We outline these latest developments there, noting here only that they relate to the great increase in *online administration and reporting* of tests, to the development of *computer programs that simulate human judgment* in the analysis of test responses and to *computer-adaptive testing*. These recent developments are revolutionizing certain aspects of the testing enterprise.

Major Forces

There is an alternative to chronology for viewing the history of testing. We can examine the major forces, trends, or recurring themes that helped create the field and brought it to the present. This is a riskier approach than the chronological one because we may overlook a significant trend or misjudge the influence of a force, hence offering easy prey to the critic. It is difficult to overlook a chronological period. However, for providing the new student with insights, this second approach may be more fruitful. Hence, we will take the risk. We identify here *six major forces* that have shaped the field of testing as we know it today.

The Scientific Impulse

The first major force influencing the development of testing is the scientific impulse. This force has prevailed throughout the history of testing. The writings of Galton, E. L. Thorndike, Cattell, Binet, and the other founders are replete with references to the need to measure scientifically. Educators, too, hoped that the development and application of "new-type" tests would make the educational enterprise scientific. The subtitle of Binet and Simon's 1905 article refers to "the necessity of establishing a scientific diagnosis." The opening sentence of Thorndike's (1904) introduction to the theory of mental measurement states that "experience has sufficiently shown that the facts of human nature can be made the material for quantitative science" (p. v). This concern for being scientific, along with the concern for scorer reliability, motivated development of early achievement tests. Finally, the field of clinical psychology, which we will treat more fully and which has been one of the primary fields of application for testing, has resolutely proclaimed its allegiance to a scientific approach. Many other professions—for example, medicine, law, and social work—have emphasized practice. Clinical psychology, however, has always maintained that it was part science and part practice, utilizing what the field calls the *scientist-practitioner model* (see Jones & Mehr, 2007).

KEY POINTS SUMMARY 1.11 Major Forces in the History of Testing

- The Scientific Impulse
- Concern for the Individual
- Practical Applications
- Statistical Methodology
- The Rise of Clinical Psychology
- Computers

Concern for the Individual

Testing has grown around an intense interest in the individual. This orientation is perhaps inevitable since testing deals with individual differences. This is part of the "differential perspective" we referenced earlier. Recall that one strand in the immediate background for establishing the roots of testing was the upsurge in concern for the welfare of the mentally ill. Many, though not all, of the practical applications mentioned later related to concern for individuals. Binet's work aimed to identify individuals who could profit more from special schools than from regular schools. Wechsler's first test aimed to give a fairer measure of intelligence for adults. The original SATs were intended to eliminate or minimize any disadvantage students from less-affluent secondary schools might have in entering college. Vocational interest measures aimed to help individualize job selection. In reading a fair selection of test manuals and the professional literature of testing throughout its history, one is struck by the frequent references to improving the lot of individuals.

Practical Applications

Virtually every major development in testing resulted from work on a practical problem. Binet tried to solve a practical problem for the Parisian schools. Wechsler wanted a better test for his adult clinical patients. The MMPI aimed to help in the diagnosis of patients at one hospital. The SAT emerged as a cooperative venture among colleges to select students from diverse secondary school experiences. Prototypes of the first group-administered intelligence tests and personality inventories developed around the need to process large numbers of military personnel in World War I. To be sure, we can find exceptions to the pattern. In some instances, notable developments resulted from theoretical considerations. However, the overall pattern seems quite clear: Testing has developed in response to practical needs. If you like the applied side of psychology, you should like the field of testing.

Statistical Methodology

The development of testing has an intriguing interactive relationship with the development of statistical methodology. One ordinarily thinks of this as a one-way street: Testing borrows methods from statistics. However, a number of statistical methods were invented specifically in response to developments in testing. The methods were then widely adopted in other fields. The first example was the display of bivariate data, invented by Galton. To further this work, Galton induced the English mathematician Karl Pearson to create the correlation coefficient. Spearman then spun off his rank-order version of the correlation. More important, in crafting his theory of intelligence, Spearman worked out the method of tetrad difference, the conceptual grandparent of modern factor analysis. The great leap forward in factor analysis came with Thurstone's work on primary mental abilities. Many of the further elaborations of factor analysis resulted from the continuing Spearman–Thurstone war of words and data over the fundamental nature of intelligence, an active battlefield to this day. Thus, the history of testing has gone hand-in-glove with the history of at least certain statistical methods.

The Rise of Clinical Psychology

Clinical psychology is one of the major areas of application for testing. In turn, clinical psychology is one of the major areas of the application of psychology. This is particularly true if we construe the term *clinical* broadly to include counseling psychology, school psychology, and the applied side of neuropsychology. On the one hand, persons in clinical practice have needed, pressed for, and helped to create a plethora of tests. On the other hand, as new tests have arisen, those in clinical practice have utilized them.

The early history of clinical psychology reads much like the early history of testing: the Binet, the Rorschach, and so on. As new tests came along, clinicians used them. In many instances, it was clinicians who developed the tests. Clinical psychologists participated actively in the military during World War I and World War II. Following World War II, the federal government invested heavily in the training of clinical psychologists. This led to explosive growth of the profession. With that growth came growth in the use of tests and the need for newer tests. This reciprocal relationship continues today. In terms of absolute numbers of tests used, the field of education leads the world of testing. In terms of the dizzying array of various types of tests available today, the clinical field (broadly conceived) has been most influential.

Computers

Computers have profoundly influenced the development of testing. As previously noted, this is a very recent phenomenon. The electronic computer was invented in 1946 and became

Courtesy of Scantron.com

FIGURE 1.7
A modern small-scale scanner, OpScan 4es OMR.

commercially available in 1951. However, mainframe computers were not in widespread use until the 1960s. Desktop models first appeared around 1980 and proliferated beginning in the mid-1980s. Thus, virtually all of the applications of computer technology to testing occurred only in the most recent historical stage we sketched earlier.

To tell the story of computers' effect on testing, we need first to distinguish between scanners and computers. There is much confusion on this point. A **scanner** is an electrical or electronic device that counts marks on a test answer sheet. The machine is sometimes called a mark-sense scanner or reader. Despite popular reference to "computer answer sheets," answer sheets are not put into a computer. They are put into a scanner (see example in Figure 1.7). The output from the scanner may (or may not) be input to a computer.

Will answer sheets and scanners become tomorrow's dinosaurs? Quite possibly. We are already seeing examinees input answers directly to a computer from a keyboard. This is being done in classrooms, clinics, human resource offices, and even in centers located in malls. Furthermore, with improvements in voice recognition, answers to test questions (delivered, for example, over phone lines) can now be spoken. The spoken words are decoded and scored by comparison with acceptable answer templates. No doubt, additional technological wonders lie just ahead.

KEY POINTS SUMMARY 1.12 The Computer-Testing Relationship

- Statistical Processing
- Score Reporting
- Test Administration

Now we move on to computers. There are three major aspects to the relationship between computers and testing. There is a historical sequence to these aspects, but they differ more in character than in temporal order. Moreover, although there is a historical sequence, it is cumulative. That is, once a phase was entered, it stayed with the field. A new phase added to rather than replaced the earlier phase.

In the *first* phase, computers simply aided in *statistical processing* for test research. This phase began almost as soon as computers became commercially available. This was a tremendous boon to testing, since it allowed for much larger research programs and routine use of sophisticated methodology. This type of development continues apace today. Desktop computers permit nearly every researcher to perform analyses in mere seconds that in the past would have required entire teams of researchers months to perform. In later exercises in this book, you will easily perform computer analyses that would have constituted a master's thesis 40 years ago.

In the *second* phase of the computer–testing relationship, computers prepared *reports of test scores*. This began with very simple reports, particularly useful for large-scale testing programs. A computer program would determine raw scores (e.g., number of right answers) using input from a scanner, convert raw scores into normed scores, and print out lists of examinee names and scores. Previously, all these operations had to be performed by hand. Computer-printed reports of this type came into widespread use in the early 1960s. By today's standards, such reports were primitive: they featured all capital letters, all the same size type, carbon paper copies, and so on. At the time, they were wonders.

The later stages of this phase evolved naturally from the earlier, simple reports. Test developers gained expertise in programming skill. They also saw the creative possibilities of computer-based reports. By the early 1970s, we began to see a profusion of ever more elaborate computer-generated reports. Printer capabilities also exploded. Graphic displays and variations in type style now accompanied numerical information, the staple of earlier reports.

A major development in this phase was the preparation of **interpretive reports**. Reports of test performance were no longer confined to numbers. Performance might now be described with simple words or even continuous narrative, as if written by a professional psychologist. We explain how such reports are prepared in Chapter 3, and we sprinkle examples of actual reports throughout later chapters.

The *third* phase of the computer–testing relationship deals with *test administration* by computer. From the test developer's perspective, there are two distinctly different types of test administration by computer. The examinee may not be aware of the difference. The first type, computer-based test administration, simply presents on a computer screen (the video monitor) the test questions very much as they appear in a printed booklet. Just put the test questions in a text file and have them pop up on the screen. The examinee inputs an answer on the keyboard. Nothing very exciting here, but it is paperless.

The second type, **computer-adaptive testing**, is revolutionary. Here, the computer not only presents the items but also selects the next item based on the examinee's previous responses. This is one of the most rapidly growing areas of testing today. Chapter 6 (pp. 181–183) provides a description of how computer-adaptive testing works.

This third phase of computer applications is now entering a new arena. In the recent past, the online completion of tests has become common. In one field, vocational interest assessment, online completion of inventories and reporting of scores are becoming standard. Other areas of application are not far behind. The main issue here is not so much the completion of the test but the delivery of test information (reports—many quite elaborate) to individuals without any training in interpreting that information and possibly without ready access to professional advice. As we will see in later chapters, interpretation of test information is not always a simple matter. Psychology has always emphasized the need for appropriate training in the interpretation of test information. Online reporting, quite apart from administration, creates a whole new scenario.

Finally, there is the emerging application of what is called **automated scoring**. This means a computer program has been developed to simulate human judgment in the scoring of such products as essays, architectural plans, and medical diagnoses. Take, for example, an essay written by a college student. Typically, it would be scored (graded) by a faculty member. In very important cases, two or three faculty members might grade it, with the grades being averaged. With automated scoring, a computer program grades the essay! Such programs, introduced just within the past few years, are now being used in place of "human raters" for scoring of essays and other such products in large-scale testing programs. We simply note here that this is a "new ball game"—one likely to see substantial growth within the next decade with ever-wider areas of application. How about, for example, the computer scoring of responses to the Rorschach inkblot test? See Bennett and Zhang (2016) and Dikli (2006) for descriptions of some of these efforts.

A Final Word on Forces

In this section, we have included only forces that affected most, if not all, types of tests. There have been other forces and trends largely restricted to one or a few types of tests. For example, cognitive psychology has affected intelligence testing. The accountability movement has affected achievement tests. We reserve treatment of these more restricted influences to the chapters on specific types of tests, for example, cognitive psychology in Chapter 8 and accountability in Chapter 11.

By Way of Definition

The final method for introducing the world of testing is by way of definition. What exactly do we mean by the term *test* or *testing*? From a strictly academic perspective, this would ordinarily be the first way to introduce a field. However, it is also a very dry, boring way to start. Thus, we preferred to use other avenues of entry. Perhaps more important, having considered the other issues treated so far, we are now in a better position to reflect on some alternative definitions and to appreciate the differences in various sources.

Identifying a consensus definition of a "test" turns out to be surprisingly difficult. The word is used in many ways and different sources, even when concentrating on psychological testing, emphasize different matters. Many of the definitions are purely circular, saying, in effect, a test is what you use when you are testing: distinctly unhelpful. Nevertheless, to guide some of our later thinking, we shall attempt to abstract from various sources what seem to be the key elements. There seem to be six common elements in what we mean by a "test" in the context of the behavioral sciences.

First, a test is some type of *procedure or device*. All agree on that point. It may be helpful for us to add that a test is a procedure or device that yields information. Perhaps that is too obvious to state, but stating it will help later discussions. Hence, we add this item as a second point: a test yields *information*. Third, the procedure or device yields information about *behavior*. This aspect of the definition is what sets a test apart from, say, physical measurements such as height and weight or medical tests such as those used to detect a viral condition. In earlier, behaviorist-oriented times, "behavior" was construed narrowly to include only externally observed behavior. In today's cognitive-oriented milieu, we construe the term more broadly to include cognitive processes. In fact, to make that explicit, we will expand the object of testing to include *behavior and cognitive processes*.

Fourth, many definitions emphasize that the test yields information only about a *sample* of behavior. When testing, we do not ordinarily take an exhaustive census of all of a person's behavior or cognitive processes. Rather, we examine a sample of the behavior or cognitive processes, often a rather small sample. This notion will become crucial in our consideration of reliability and validity. Fifth, a test is a *systematic, standardized* procedure. This is one of the most distinctive characteristics of a test. This characteristic differentiates a test from such sources of information as an informal interview or anecdotal observations—both of which may be useful sources of information, but are not tests.

We need to digress at this point to clarify a potentially confusing matter of terminology. There are *three* uses of the term **standardized** in the world of testing. First, when that term is used in the definition of testing, it refers to uniform procedures for administering and scoring. There are definite, clearly specified methods for administering the test, and there are rules for scoring the test. It is essential that the test be administered and scored according to these procedures. Second, in other contexts, standardized means that the test has norms—for example, national norms based on thousands of cases. In fact, the process of collecting these normative data is often referred to as the standardization program for the test. Clearly, this is a different meaning of the term *standardized* than the first meaning. One can have a test with fixed directions and scoring procedures

without having any type of norms for the test. A third meaning, encountered especially in media reporting and public discussions, equates standardized testing with group-administered, machine-scored, multiple-choice tests of ability and achievement. For example, a newspaper headline may report "Local students improve on standardized tests" or "Cheating on standardized tests alleged." Or a friend may say, referring to performance on the SAT or ACT college admissions test, "I don't do very well on standardized tests." This third meaning is obviously much more restrictive than either of the first two. It is important for the student of psychological testing to distinguish among these three meanings of the term *standardized*.

A sixth and final element in the various definitions is some reference to *quantification or measurement*. That is, we finally put the information in numerical form. This element is quite explicit in some sources and seems to be implied in the others. The quantification may occur in a very crude form or in a highly sophisticated manner. For example, a crude quantification may involve forming two groups (depressed and nondepressed or competent and not competent). A more sophisticated kind of measurement may involve a careful scaling akin to measurement of height or weight.

Various sources differ on one matter of the definition of "test," namely the extent to which testing is evaluative. Some of the definitions stop with the information; other definitions include reference to an evaluative dimension, an inference or conclusion drawn from the information. Some books handle this point by distinguishing among the terms *test*, *assessment*, and *evaluation*. For example, some authors imply differences among these three statements: We *tested* Abigail's intelligence. We *assessed* Abigail's intelligence. We *evaluated* Abigail's intelligence. In many sources, the three terms are used interchangeably. For example, the *Standards* (AERA et al., 2014) seem to merge the terms, defining a "test" as "an evaluative device" (p. 224) and "psychological testing" as "The use of tests . . . to assess" (p. 222). We do not take a hard-and-fast position on this matter. We simply note that various sources handle the matter differently.

From the foregoing discussion, we formulate the following definition: **A test is a standardized process or device that yields information about a sample of behavior or cognitive processes in a quantified manner**.

KEY POINTS SUMMARY 1.13 Elements in the Definition of "Test"

- Process or Device
- Yields Information
- Behavior or Cognitive Processes
- Sample of . . .
- Standardized
- Quantified

SUMMARY

1. We classify tests into five major categories: mental ability, achievement, personality, interests, and neuropsychological tests, with several subdivisions within some of the major categories. Use the acronym MAPIN to remember these categories.
2. Tests may also be characterized according to whether they are (a) paper-and-pencil or performance, (b) speed or power, (c) individually or group administered, (d) dependent on maximum or typical performance, and (e) norm-referenced or criterion-referenced in their interpretation.
3. The principal uses of tests include clinical, educational, personnel, and research.
4. Four important assumptions undergird the testing enterprise:
   - That people have traits and that differences among people in these traits are important.
   - That we can quantify these traits.
   - That the traits have some reasonable degree of stability.
   - That our quantification of the traits bears some relationship to actual behavior.
5. The four fundamental questions in testing relate to:
   - Reliability—the stability of a measure.
   - Validity—what a test really measures.
   - Norms—the framework for interpreting test scores.
   - Fairness—measuring equivalently for different groups.

We study these topics in depth in Chapters 3, 4, 5, and 7. How tests are developed, covered in Chapter 6, and practical concerns such as time and cost are also important considerations.

6. We identified seven major periods in the history of testing. Understanding the dominant themes of these periods provides perspective on current issues in testing. The periods and the titles we gave them are:

• Up to 1840	Remote Background
• 1840–1880	Setting the Stage
• 1880–1915	The Roots
• 1915–1940	Flowering
• 1940–1965	Consolidation
• 1965–2000	Just Yesterday
• 2000–Present	And Now

7. We identified six major forces influencing the development of testing as the field currently exists: the scientific impulse, concern for the individual, practical applications, statistical methodology, the rise of clinical psychology, and computers.
8. We developed the following six-element definition of a test: *A test is a standardized process or device that yields information about a sample of behavior or cognitive processes in a quantified manner*.

KEY TERMS

achievement tests
Alfred Binet
automated scoring
Charles Spearman
classical test theory
cognitive ability tests
computer-adaptive testing
criterion-referenced interpretation
differential perspective
evidence-based practice
fairness
Francis Galton
group test
individual test
interpretive reports
item response theory
James McKeen Cattell
maximum performance
neuropsychological tests
norm-referenced interpretation
norms
objective personality tests
paper-and-pencil test
performance test
power test
projective techniques
reliability
scanner
speed (or speeded) test
standardized
typical performance
validity
vocational interest measures

EXERCISES

1. Through your university library, access Cattell's 1890 article in which he coined the term "mental test." See References in this book for the full citation for the article. Or you can access the article at http://psychclassics.yorku.ca. Find the article and check the list of tests described therein. What do you think of these tests as predictors of college success?
2. If you happen to be taking a history course at the same time as you are using this text, try to relate something in your history course to the periods of development in the history of testing. Do you observe any trends or forces in your history course that might have influenced developments in testing?
3. Most universities and even many departments within larger universities have their own scanners for processing test answer sheets. See if you can locate a scanner and watch it in operation. What is the "output" from the scanner?
4. Think of tests you have taken other than classroom tests. Categorize each test according to these distinctions: (a) paper-and-pencil versus performance, (b) speed versus power, (c) individually or group administered, (d) dependent on maximum or typical performance, and (e) norm-referenced or criterion-referenced in interpretation.
5. Use the website http://psychclassics.yorku.ca/ to access Alfred Binet's classic work *New Methods for the Diagnosis of the Intellectual Level of Subnormals*, written in 1905. (Notice the use of terms such as imbecile and idiot. These would be considered derogatory terms today but were standard, clinical descriptors at that time.) From reading just the first few paragraphs of Binet's work, what do you think he was trying to do?

6. Access this website: https://nces.ed.gov/nationsreportcard/ for results of the National Assessment of Educational Progress (NAEP). How many grades are tested by NAEP? For how many school subjects are reports available? Access the report for one subject of interest to you. What are some of the major findings for that subject?

7. Here are three traits: height, intelligence, friendliness. In which of these traits do you think people differ the most?

8. Recall our observation that many tests are known mainly by their initials. See if you can remember the full names for each of these sets of initials.

   EDI

   SII

   LSAT

   BDI-II

9. Many of the classical documents in the history of testing (e.g., those of Darwin, Galton, Cattell, and Binet) can be viewed at this website: http://psychclassics.yorku.ca/. Check it out. Skim some of the documents to get a flavor for how the authors approached their topics.

10. To see an interesting graphical presentation of the relationships among people working on early intelligence tests, check this website: http://www.intelltheory.com/. Click on Interactive Map. How does Piaget fit in the map? How about Anastasi? You can access brief biographies of most of the people mentioned in our description of the history of testing by clicking on a name in the interactive map.