CHAPTER 1

The History of Psychological Testing

TOPIC 1A The Origins of Psychological Testing

The history of psychological testing is a fascinating story and has abundant relevance to present-day practices. After all, contemporary tests did not spring from a vacuum; they evolved slowly from a host of precursors introduced over the last one hundred years. Accordingly, Chapter 1 features a review of the historical roots of present-day psychological tests. In Topic 1A, The Origins of Psychological Testing, we focus largely on the efforts of European psychologists to measure intelligence during the late nineteenth century and pre–World War I era. These early intelligence tests and their successors often exerted powerful effects on the examinees who took them, so the first topic also incorporates a brief digression documenting the pervasive importance of psychological test results. Topic 1B, Early Testing in the United States, catalogues the profusion of tests developed by American psychologists in the first half of the twentieth century.

Psychological testing in its modern form originated little more than one hundred years ago in laboratory studies of sensory discrimination, motor skills, and reaction time. The British genius Francis

Galton (1822–1911) invented the first battery of tests, a peculiar assortment of sensory and motor measures, which we review in the following. The American psychologist James McKeen Cattell (1860–1944) studied with Galton and then, in 1890, proclaimed the modern testing agenda in his classic paper entitled "Mental Tests and Measurements." He was tentative and modest when describing the purposes and applications of his instruments:

> Psychology cannot attain the certainty and exactness of the physical sciences, unless it rests on a foundation of experiment and measurement. A step in this direction could be made by applying a series of mental tests and measurements to a large number of individuals. The results would be of considerable scientific value in discovering the constancy of mental processes, their interdependence, and their variation under different circumstances. Individuals, besides, would find their tests interesting, and, perhaps, useful in regard to training, mode of life or indication of disease. The scientific and practical value of such tests would be much increased should a uniform system be adopted, so that determinations made at different times and places could be compared and combined. (Cattell, 1890)

Cattell's conjecture that "perhaps" tests would be useful in "training, mode of life or indication of disease" must certainly rank as one of the prophetic understatements of all time. Anyone reared in the Western world knows that psychological testing has emerged from its timid beginnings to become a big business and a cultural institution that permeates modern society. To cite just one example, consider the number of standardized achievement and ability tests administered in the school systems of the United States. Although it is difficult to obtain exact data on the extent of such testing, an estimate of 200 million per year is probably not extreme (Medina & Neill, 1990). Of course, the total number of tests administered yearly also includes millions of personality tests and untold numbers of the thousands of other kinds of tests now in existence (Conoley & Kramer, 1989, 1992; Mitchell, 1985; Sweetland & Keyser, 1987). There is no doubt that testing is pervasive. But does it make a difference?

THE IMPORTANCE OF TESTING

Tests are used in almost every nation on earth for counseling, selection, and placement. Testing occurs in settings as diverse as schools, civil service, industry, medical clinics, and counseling centers. Most persons have taken dozens of tests and thought nothing of it. Yet, by the time the typical individual reaches retirement age, it is likely that psychological test results will help shape his or her destiny. The deflection of the life course by psychological test results might be subtle, such as when a prospective mathematician qualifies for an accelerated calculus course based on tenth-grade achievement scores. More commonly, psychological test results alter individual destiny in profound ways. Whether a person is admitted to one college and not another, offered one job but refused a second, diagnosed as depressed or not—all such determinations rest, at least in part, on the meaning of test results as interpreted by persons in authority. Put simply, psychological test results change lives. For this reason it is prudent—indeed, almost mandatory—that students of psychology learn about the contemporary uses and occasional abuses of testing. In Case Exhibit 1.1, the life-altering aftermath of psychological testing is illustrated by means of several true case history examples.

The importance of testing is also evident from historical review. Students of psychology generally regard historical issues as dull, dry, and pedantic, and sometimes these prejudices are well deserved. After all, many textbooks fail to explain the relevance of historical matters and provide only vague sketches of early developments in mental testing. As a result, students of psychology often conclude incorrectly that historical issues are boring and irrelevant.

In reality, the history of psychological testing is a captivating story that has substantial relevance to present-day practices. Historical developments are pertinent to contemporary testing for the following reasons:

1. A review of the origins of psychological testing helps explain current practices that might other-

CASE EXHIBIT 1.1

THE CONSEQUENCES OF TEST RESULTS

The importance of psychological testing is best illustrated by example. Consider these brief vignettes:

- A shy, withdrawn 7-year-old girl is administered an IQ test by a school psychologist. Her score is phenomenally higher than the teacher expected. The student is admitted to a gifted and talented program where she blossoms into a self-confident and gregarious scholar.
- Three children in a family living near a lead smelter are exposed to the toxic effects of lead dust and suffer neurological damage. Based in part on psychological test results that demonstrate impaired intelligence and shortened attention span in the children, the family receives an $8 million settlement from the company that owns the smelter.
- A candidate for a position as police officer is administered a personality inventory as part of the selection process. The test indicates that the candidate tends to act before thinking and resists supervision from authority figures. Even though he has excellent training and impresses the interviewers, the candidate does not receive a job offer.
- A student, unsure of what career to pursue, takes a vocational interest inventory. The test indicates that she would like the work of a pharmacist. She signs up for a prepharmacy curriculum but finds the classes to be both difficult and boring. After three years, she abandons pharmacy for a major in dance, frustrated that she still faces three more years of college to earn a degree.
- An applicant to graduate school in clinical psychology takes the Minnesota Multiphasic Personality Inventory (MMPI). His recommendations and grade point average are superlative, yet he must clear the final hurdle posed by the MMPI. His results are reasonably normal but slightly defensive; by a narrow vote, the admissions committee extends him an invitation. Ironically, this is the only graduate school to admit him—nineteen others turn him down. He accepts the invitation and becomes enchanted with the study of psychological assessment. Many years later, he writes this book.

wise seem arbitrary or even peculiar. For example, why do many current intelligence tests incorporate a seemingly nonintellective capacity, namely, short-term memory for digits? The answer is, in part, historical inertia—intelligence tests have always included a measure of digit span.

2. The strengths and limitations of testing also stand out better when tests are viewed in historical context. The reader will discover, for example, that modern intelligence tests are exceptionally good at predicting school failure—precisely because this was the original and sole purpose of the first such instrument developed in Paris, France, at the turn of the twentieth century.
3. Finally, the history of psychological testing contains some sad and regrettable episodes that help remind us not to be overly zealous in our modern-day applications of testing. For example, based on the misguided and prejudicial

application of intelligence test results, several prominent psychologists helped ensure passage of the Immigration Restriction Act of 1924.

In later chapters, we examine the principles of psychological testing, investigate applications in specific fields (e.g., personality, intelligence, neuropsychology), and reflect on the social and legal consequences of testing. However, the reader will find these topics more comprehensible when viewed in historical context. So, for now, we begin at the beginning by reviewing rudimentary forms of testing that existed over four thousand years ago in imperial China.

RUDIMENTARY FORMS OF TESTING IN CHINA IN 2200 B.C.

Although the widespread use of psychological testing is largely a phenomenon of the twentieth century, historians note that rudimentary forms of testing date back to at least 2200 B.C. when the Chinese emperor had his officials examined every third year to determine their fitness for office (Bowman, 1989; Chaffee, 1985; DuBois, 1970; Franke, 1963; Lai, 1970; Teng, 1942–43). Such testing was modified and refined over the centuries until written exams were introduced in the Han dynasty (202 B.C.–A.D. 200). Five topics were tested: civil law, military affairs, agriculture, revenue, and geography.

The Chinese examination system took its final form about 1370 when proficiency in the Confucian classics was emphasized. In the preliminary examination, candidates were required to spend a day and a night in a small isolated booth, composing essays on assigned topics and writing a poem. The 1 to 7 percent who passed moved up to the district examinations, which required three separate sessions of three days and three nights.

The district examinations were obviously grueling and rigorous, but this was not the final level. The 1 to 10 percent who passed were allowed the privilege of going to Peking for the final round of examinations. Perhaps 3 percent of this final group passed and became mandarins, eligible for public office.

Although the Chinese developed the external trappings of a comprehensive civil service examination program, the similarities between their traditions and current testing practices are, in the main, superficial. Not only were their testing practices unnecessarily grueling, the Chinese also failed to validate their selection procedures. Nonetheless, it does appear that the examination program incorporated relevant selection criteria. For example, in the written exams beauty of penmanship was weighted very heavily. Given the highly stylistic features of Chinese written forms, good penmanship was no doubt essential for clear, exact communication. Thus, penmanship was probably a relevant predictor of suitability for civil service employment. In response to widespread discontent, the examination system was abolished by royal decree in 1906 (Franke, 1963).

PSYCHIATRIC ANTECEDENTS OF PSYCHOLOGICAL TESTING

Most historians trace the beginnings of psychological testing to the experimental investigation of individual differences that flourished in Germany and Great Britain in the late 1800s. There is no doubt that early experimentalists such as Wilhelm Wundt, Francis Galton, and James McKeen Cattell laid the foundations for modern-day testing, and we will review their contributions in detail. But psychological testing owes as much to early psychiatry as it does to the laboratories of experimental psychology. In fact, the examination of the mentally ill around the middle of the nineteenth century resulted in the development of numerous early tests (Bondy, 1974). These early tests featured the absence of standardization and were consequently relegated to oblivion. They were nonetheless influential in determining the course of psychological testing, so it is important to mention a few typical developments from this era.

In 1885, the German physician Hubert von Grashey developed the antecedent of the memory drum as a means of testing brain-injured patients. His subjects were shown words, symbols, or pictures through a slot in a sheet of paper that was moving slowly over the stimuli. Grashey found that many patients could recognize stimuli in their totality but could not identify them when shown

through the moving slot. Shortly thereafter, the German psychiatrist Conrad Rieger developed an excessively ambitious test battery for brain damage. His battery took over 100 hours to administer and soon fell out of favor.

In summary, early psychiatry contributed to the mental test movement by showing that standardized procedures could help reveal the nature and extent of symptoms in the mentally ill and brain-injured patients. Most of the early tests developed by psychiatrists faded into oblivion, but a few procedures were standardized and perpetuate themselves in modern variations (Bondy, 1974).

THE BRASS INSTRUMENTS ERA OF TESTING

Experimental psychology flourished in the late 1800s in continental Europe and Great Britain. For the first time in history, psychologists departed from the wholly subjective and introspective methods that had been so fruitlessly pursued in the preceding centuries. Human abilities were instead tested in laboratories. Researchers used objective procedures that were capable of replication. Gone were the days when rival laboratories would have raging arguments about "imageless thought," one group saying it existed, another group saying that such a mental event was impossible.

Even though the new emphasis on objective methods and measurable quantities was a vast improvement over the largely sterile mentalism that preceded it, the new experimental psychology was itself a dead end, at least as far as psychological testing was concerned. The problem was that the early experimental psychologists mistook simple sensory processes for intelligence. They used assorted brass instruments to measure sensory thresholds and reaction times, thinking that such abilities were at the heart of intelligence. Hence, this period is sometimes referred to as the Brass Instruments era of psychological testing.

In spite of the false start made by early experimentalists, at least they provided psychology with an appropriate methodology. Such pioneers as Wundt, Galton, Cattell, and Clark Wissler showed that it was possible to expose the mind to scientific scrutiny and measurement. This was a fateful change in the axiomatic assumptions of psychology, a change that has stayed with us to the current day.

Most sources credit Wilhelm Wundt (1832–1920) with founding the first psychological laboratory in 1879 in Leipzig, Germany. It is less well recognized that he was measuring mental processes years before, at least as early as 1862, when he experimented with his thought meter (Diamond, 1980). This device was a calibrated pendulum with needles sticking off from each side. The pendulum would swing back and forth, striking bells with the needles. The observer's task was to take note of the position of the pendulum when the bells sounded. Of course, Wundt could adjust the needles beforehand and thereby know the precise position of the pendulum when each bell was struck. Wundt thought that the difference between the observed pendulum position and the actual position would provide a means of determining the swiftness of thought of the observer.

Wundt's analysis was relevant to a longstanding problem in astronomy. The problem was that two or more astronomers simultaneously using the same telescope (with multiple eyepieces) would report different crossing times as the stars moved across a grid line on the telescope. Even in Wundt's time, it was a well-known event in the history of science that Kinnebrook, an assistant at the Royal Observatory in England, had been dismissed in 1796 because his stellar crossing times were nearly a full second too slow (Boring, 1950). Wundt's analysis offered another explanation that did not assume incompetence on the part of anyone. Put simply, Wundt believed that the speed of thought might differ from one person to the next:

> For each person there must be a certain speed of thinking, which he can never exceed with his given mental constitution. But just as one steam engine can go faster than another, so this speed of thought will probably not be the same in all persons. (Wundt, 1862, as translated in Rieber, 1980)

This analysis of telescope reporting times seems simplistic by present-day standards and overlooks the possible contribution of such factors as attention,

motivation, and self-correcting feedback from prior trials. On the positive side, this was at least an empirical analysis that sought to explain individual differences instead of trying to explain them away. And that is the relevance to current practices in psychological testing. However crudely, Wundt measured mental processes and begrudgingly acknowledged individual differences.[1]

Galton and the First Battery of Mental Tests

Sir Francis Galton (1822–1911) pioneered the new experimental psychology in nineteenth-century Great Britain. Galton was obsessed with measurement, and his intellectual career seems to have been dominated by a belief that virtually anything was measurable. His attempts to measure intellect by means of reaction time and sensory discrimination tasks are well known. Yet, to appreciate his wide-ranging interests, the reader should be apprised that Galton also devised techniques for measuring beauty, personality, the boringness of lectures, and the efficacy of prayer, to name but a few of the endeavors that his biographer has catalogued in elaborate detail (Pearson 1914, 1924, 1930ab).

Galton was a genius who was more interested in the problems of human evolution than in psychology per se (Boring, 1950). His two most influential works were *Hereditary Genius* (1869), an empirical analysis purporting to prove that genetic factors were overwhelmingly important for the attainment of eminence, and *Inquiries into Human Faculty and Its Development* (1883), a disparate series of essays that emphasized individual differences in mental faculties.

Boring (1950) regards *Inquiries* as the beginning of the mental test movement and the advent of the scientific psychology of individual differences. The book is a curious mixture of empirical research and speculative essays on topics as diverse as "just perceptible differences" in lifted weight and diminished fertility among inbred animals. There is, nonetheless, a common theme uniting these diverse essays; Galton demonstrates time and again that individual differences not only exist but are objectively measurable.

Galton borrowed the time-consuming psychophysical procedures practiced by Wundt and others on the European continent and adapted them to a series of simple and quick sensorimotor measures. Thus, he continued the tradition of brass instruments mental testing but with an important difference: his procedures were much more amenable to the timely collection of data from hundreds if not thousands of subjects. Because of his efforts in devising practicable measures of individual differences, historians of psychological testing usually regard Galton as the father of mental testing (Goodenough, 1949; Boring, 1950).

To further his study of individual differences, Galton set up a psychometric laboratory in London at the International Health Exhibition in 1884. It was later transferred to the London Museum, where it was maintained for six years. Various anthropometric and psychometric measures were arranged on a long table at one side of a narrow room. Subjects were admitted at one end for threepence and given successive tests as they moved down the table. At least 17,000 individuals were tested during the 1880s and 1890s. About 7,500 of the individual data records have survived to the present day (Johnson et al., 1985).

The tests and measures involved both the physical and behavioral domains. Physical characteristics assessed were height, weight, head length, head breadth, arm span, length of middle finger, and length of lower arm, among others. The behavioral tests included strength of hand squeeze determined by dynamometer, vital capacity of the lungs measured by spirometer, visual acuity, highest audible tone, speed of blow, and reaction time (RT) to both visual and auditory stimuli.

Ultimately, Galton's simplistic attempts to gauge intellect with measures of reaction time and sensory discrimination proved fruitless. Nonetheless, he did provide a tremendous impetus to the testing movement by demonstrating that objective tests could be devised and that meaningful scores could be obtained through standardized procedures.

1. This emphasis upon individual differences was rare for Wundt. He is more renowned for proposing common laws of thought for the average adult mind.

Cattell Imports Brass Instruments to the United States

James McKeen Cattell (1860–1944) studied the new experimental psychology with both Wundt and Galton before settling at Columbia University where, for twenty-six years, he was the undisputed dean of American psychology. With Wundt, he did a series of painstakingly elaborate RT studies (1880–1882), measuring with great precision the fractions of a second presumably required for different mental reactions. He also noted, almost in passing, that he and another colleague had small but consistent differences in RT. Cattell proposed to Wundt that such individual differences ought to be studied systematically. Although Wundt acknowledged individual differences, he was philosophically more inclined to study general features of the mind, and he offered no support for Cattell's proposal (Fancher, 1985).

But Cattell received enthusiastic support for his study of individual differences from Galton, who had just opened his psychometric laboratory in London. After corresponding with Galton for a few years, Cattell arranged for a two-year fellowship at Cambridge so that he could continue the study of individual differences. Cattell opened his own research laboratory and developed a series of tests that were mainly extensions and additions to Galton's battery.

Cattell (1890) invented the term *mental test* in his famous paper entitled "Mental Tests and Measurements." This paper described his research program, detailing ten mental tests he proposed for use with the general public. These tests were clearly a reworking and embellishment of the Galtonian tradition:

Strength of hand squeeze as measured by dynamometer
Rate of hand movement through a distance of 50 centimeters
Two-point threshold for touch—minimum distance at which two points are still perceived as separate
Degree of pressure needed to cause pain—rubber tip pressed against the forehead
Weight differentiation—discern the relative weights of identical-looking boxes varying by one gram from 100 to 110 grams
Reaction time for sound—using a device similar to Galton's
Time for naming colors
Bisection of a 50-centimeter line
Judgment of 10 seconds of time
Number of letters repeated on one hearing

Strength of hand squeeze seems a curious addition to a battery of mental tests, a point that Cattell (1890) addressed directly in his paper. He was of the opinion that it was impossible to separate bodily energy from mental energy. Thus, in Cattell's view, an ostensibly physiological measure such as dynamometer pressure was an index of one's mental power as well. Clearly, the physiological and sensory bias of the entire test battery reflects its strongly Galtonian heritage (Fancher, 1985).

In 1891, Cattell accepted a position at Columbia University, at that time the largest university in the United States. His subsequent influence on American psychology was far in excess of his individual scientific output and was expressed in large part through his numerous and influential students (Boring, 1950). Among his many famous doctoral students and the years of their degrees were E. L. Thorndike (1898) who made monumental contributions to learning theory and educational psychology; R. S. Woodworth (1899) who was to author the very popular and influential *Experimental Psychology* (1938); and E. K. Strong (1911) whose Vocational Interest Blank—since revised—is still in wide use. But among Cattell's students, it was probably Clark Wissler (1901) who had the greatest influence on the early history of psychological testing.

Wissler obtained both mental test scores and academic grades from more than 300 students at Columbia University and Barnard College. His goal was to demonstrate that the test results could predict academic performance. With our early twenty-first-century perspective on research and testing, it seems amazing that the early experimentalists waited so long to do such basic validational research. Wissler's (1901) results showed virtually no tendency for the mental test scores to correlate with academic achievement. For example, class standing correlated .16 with memory for number lists, –.08

with dynamometer strength, .02 with color naming, and –.02 with reaction time. The highest correlation (.16) was statistically significant because of the large sample size. However, so humble a correlation carries with it very little predictive utility.[2]

Also damaging to the brass instruments testing movement was the very modest correlations between the mental tests themselves. For example, color naming and hand movement speed correlated only .19, while RT and color naming correlated –.15. Several physical measures such as head size (a holdover measure from the Galton era) were, not surprisingly, also uncorrelated with the various sensory and RT measures.

With the publication of Wissler's (1901) discouraging results, experimental psychologists largely abandoned the use of RT and sensory discrimination as measures of intelligence. From one standpoint, this turning away from the brass instruments approach was a desirable development in the history of psychological testing. The way was thereby paved for immediate acceptance of Alfred Binet's more sensible and useful measures of higher mental processes.

But in other respects, the abandonment of RT and sensory measures was premature and unfortunate. After all, by contemporary standards Wissler's research methods revealed an extraordinary psychometric naivete. By using only bright college students as subjects, Wissler had inadvertently introduced an extreme restriction of range, which would invariably reduce the size of his correlations. If a more heterogeneous sample of subjects had been used, the correlations would have been substantially larger. In addition, certain measures such as RT were inherently unreliable because of the small number of trials per subject. Such unreliability in a measure also places a severe restriction on the upper bounds of correlation coefficients.

If Wissler's (1901) negative findings had been more skeptically scrutinized, it might not have been a full 70 years later until RT was resurrected as a potentially useful intellectual measure. Correlations of –.40 between complex forms of RT and intelligence are not at all uncommon (Jensen, 1982).[3]

But that is getting ahead of the story. The more common reaction among psychologists in the early 1900s was to begrudgingly conclude that Galton had been wrong in attempting to infer complex abilities from simple ones. Goodenough (1949) has likened Galton's approach to "inferring the nature of genius from the nature of stupidity or the qualities of water from those of the hydrogen and oxygen of which it is composed." The academic psychologists apparently agreed with her, and American attempts to develop intelligence tests virtually ceased at the turn of the twentieth century. For his own part, Wissler was apparently so discouraged by his results that he immediately switched to anthropology, where he became a strong environmentalist in explaining differences between ethnic groups.

The void created by the abandonment of the Galtonian tradition did not last for long. In Europe, Alfred Binet was on the verge of a major breakthrough in intelligence testing. Binet introduced his scale of intelligence in 1905, and shortly thereafter H. H. Goddard imported it to the United States, where it was applied in a manner that Gould (1981) has described as "the dismantling of Binet's intentions in America." Whether early twentieth-century American psychologists subverted Binet's intentions is an important question that we review in the next topic. First, we examine the social changes in nineteenth-century Europe that created the necessity for practical intelligence tests.

CHANGING CONCEPTIONS OF MENTAL RETARDATION IN THE 1800S

Many great inventions have been developed in response to the practical needs created by changes in

2. We discuss the correlation coefficient in more detail in Topic 3B, Concepts of Reliability. By way of quick preview, correlations can range from –1.0 to +1.0. Values near zero indicate a weak, negligible linear relationship between the two variables. For example, correlations between –.20 and +.20 are generally of minimal value for purposes of individual prediction. Note also that negative correlations indicate an inverse relationship.

3. The correlations are negative because *low* scores on RT are associated with high scores on intelligence tests.

societal values. Such is the case with intelligence tests. To be specific, the first such tests were developed by Binet in the early 1900s to help identify children in the Paris school system who were unlikely to profit from ordinary instruction. Prior to this time, there was little interest in the educational needs of children with mental retardation. A new humanism toward those with mental retardation thus created the practical problem—identifying those with special needs—that Binet's tests were to solve.

The Western world of the late 1800s was just emerging from centuries of indifference and hostility toward the psychiatrically and mentally impaired. Medical practitioners were just beginning to acknowledge a distinction between individuals with emotional disablities and mental retardation. For centuries, all such social outcasts were given similar treatment. In the Middle Ages, they were occasionally "diagnosed" as witches and put to death by burning. Later on, they were alternately ignored, persecuted, or tortured. In his comprehensive history of psychotherapy and psychoanalysis, Bromberg (1959) has an especially graphic chapter on the various forms of maltreatment toward those with mental and emotional disabilities, from which only one example will be provided here. In 1698, a prominent physician wrote a gruesome book, *Flagellum Salutis,* in which beatings were advocated as treatment "in melancholia; in frenzy; in paralysis; in epilepsy; in facial expression of feebleminded" (Bromberg, 1959).

By the early 1800s, saner minds began to prevail. Medical practitioners realized that some of those with psychiatric impairment had reversible illnesses that did not necessarily imply diminished intellect, whereas other exceptional persons, those with mental retardation, showed a greater developmental continuity and invariably had impaired intellect. In addition, a newfound humanism began to influence social practices toward individuals with psychological and mental disabilities. With this humanism there arose a greater interest in the diagnosis and remediation of mental retardation. At the forefront of these developments were two French physicians, J. E. D. Esquirol and O. E. Seguin, each of whom revolutionized thinking about those with mental retardation, thereby helping to create the necessity for Binet's tests.

Esquirol and Diagnosis in Mental Retardation

Around the beginning of the nineteenth century, many physicians had begun to perceive the difference between mental retardation (then called *idiocy*) and mental illness (often referred to as *dementia*). J. E. D. Esquirol (1772–1840) was the first to formalize the difference in writing. His diagnostic breakthrough was noting that mental retardation was a lifelong developmental phenomenon whereas mental illness usually had a more abrupt onset in adulthood. He thought that mental retardation was incurable, whereas mental illness might show improvement (Esquirol, 1845/1838).

Esquirol placed great emphasis upon language skills in the diagnosis of mental retardation. This may offer a partial explanation as to why Binet's later tests and the modern-day descendents from them are so heavily loaded on linguistic abilities. After all, the original use of the Binet scales was, in the main, to identify children with mental retardation who would not likely profit from ordinary schooling.

Esquirol also proposed the first classification system in mental retardation and it should be no surprise that language skills were the main diagnostic criteria. He recognized three levels of mental retardation: (1) those using short phrases, (2) those using only monosyllables, and (3) those with cries only, no speech. Apparently, Esquirol did not recognize what we would now call *mild mental retardation,* instead providing criteria for the equivalents of the modern-day classifications of moderate, severe, and profound mental retardation.

Seguin and Education of Individuals with Mental Retardation

Perhaps more than any other pioneer in the field of mental retardation, O. Edouard Seguin (1812–1880) helped establish a new humanism toward those with mental retardation in the late 1800s. He had been a student of Esquirol and had also studied with J. M. G. Itard (1774–1838), who is well known

for his five-year attempt to train the Wild Boy of Aveyron, a feral child who had lived in the woods for his first 11 or 12 years (Itard, 1932/1801).

Seguin borrowed from techniques used by Itard and devoted his life to developing educational programs for persons with mental retardation. As early as 1838, he had established an experimental class for such individuals. His treatment efforts earned him international acclaim and he eventually came to the United States to continue his work. In 1866, he published *Idiocy, and Its Treatment by the Physiological Method,* the first major textbook on the treatment of mental retardation. This book advocated a surprisingly modern approach to education of individuals with mental retardation and even touched on what would now be called *behavior modification.*

Such was the social and historical background that allowed intelligence tests to flourish. We turn now to the invention of the modern-day intelligence test by Alfred Binet. We begin with a discussion of the early influences that shaped his famous test.

INFLUENCE OF BINET'S EARLY RESEARCH UPON HIS TEST

As most every student of psychology knows, Alfred Binet (1857–1911) invented the first modern intelligence test in 1905. What is less well known, but equally important for those who seek an understanding of his contributions to modern psychology, is that Binet was a prolific researcher and author long before he turned his attentions to intelligence testing. The character of his early research had a material bearing on the subsequent form of his well-known intelligence test. For those who seek a full understanding of his pathbreaking influence, brief mention of Binet's early career is mandatory. For more details the reader can consult DuBois (1970), Fancher (1985), Goodenough (1949), Gould (1981), and Wolf (1973).

Binet began his career in medicine, but was forced to drop out because of a complete emotional breakdown. He switched to psychology, where he studied the two-point threshold and dabbled in the associationist psychology of John Stuart Mill (1806–1873). Later, he selected an apprenticeship with the neurologist J. M. Charcot (1825–1893) at the famous Salpetriere Hospital. Thus, for a brief time Binet's professional path paralleled that of Sigmund Freud, who also studied hysteria under Charcot. At the Salpetriere Hospital, Binet coauthored (with C. Fere) four studies supposedly demonstrating that reversing the polarity of a magnet could induce complete mood changes (e.g., from happy to sad) or transfer of hysterical paralysis (e.g., from left to right side) in a single hypnotized subject. In response to public criticism from other psychologists, Binet later published a recantation of his findings. This was a painful episode for Binet, and it sent his career into a temporary detour. Nonetheless, he learned two things through his embarrassment. First, he never again used sloppy experimental procedures that allowed for unintentional suggestion to influence his results. Second, he became skeptical of the zeitgeist (spirit of the times) in experimental psychology. Both of these lessons were applied when he later developed his intelligence scales.

In 1891, Binet went to work at the Sorbonne as an unpaid assistant and began a series of studies and publications that were to define his new "individual psychology" and ultimately to culminate in his intelligence tests. Binet was an ardent experimentalist, often using his two daughters to try out existing and new tests of intelligence. Early on, he flirted with a Cattellian approach to intelligence testing, using the standard measures of reaction time and sensory acuity on his two daughters. The results were annoyingly inconsistent and difficult to interpret. As might be expected, he found that the reaction times of his children were, on average, much slower than for adults. But on some trials his daughters' performance approached or exceeded adult levels. From these findings, Binet concluded that attention was a key component of intelligence, which was itself a very multifaceted entity. Furthermore, he became increasingly disenchanted with the brass instruments approach to measuring intelligence, which probably explains his subsequent use of measures of higher mental processes.

In addition, Binet's sensory-perceptual experiments with his children greatly influenced his views on proper testing procedures:

> The experimenter is obliged, to a point, to adjust his method to the subject he is addressing. There are certain rules to follow when one experiments on a child, just as there are certain rules for adults, for hysterics, and for the insane. These rules are not written down anywhere; each one learns them for himself and is repaid in great measure. By making an error and later accounting for the cause, one learns not to make the mistake a second time. In regard to children, it is necessary to be suspicious of two principal causes of error: suggestion and failure of attention. This is not the time to speak on the first point. As for the second, failure of attention, it is so important that it is always necessary to suspect it when one obtains a negative result. One must then suspend the experiments and take them up at a more favorable moment, restarting them 10 times, 20 times, with great patience. Children, in fact, are often little disposed to pay attention to experiments which are not entertaining, and it is useless to hope that one can make them more attentive by threatening them with punishment. By particular tricks, however, one can sometimes give the experiment a certain appeal. (Binet, 1895, quoted in Pollack, 1971)

It is interesting to contrast modern-day testing practices—which go so far as to specify the exact wording the examiner should use—with Binet's advice to exercise nearly endless patience and use entertaining tricks when testing children.

BINET AND TESTING FOR HIGHER MENTAL PROCESSES

In 1896, Binet and his Sorbonne assistant, Victor Henri, published a pivotal review of German and American work on individual differences. In this historically important paper, they argued that intelligence could be better measured by means of the higher psychological processes rather than the elementary sensory processes such as reaction time. After several false starts, Binet and Simon eventually settled on the straightforward format of their 1905 scales, discussed subsequently.

The character of the 1905 scale owed much to a prior test developed by Dr. Blin (1902) and his pupil, M. Damaye. They had attempted to improve the diagnosis of mental retardation by using a battery of assessments in 20 areas such as spoken language; knowledge of parts of the body; obedience to simple commands; naming common objects; and ability to read, write, and do simple arithmetic. Binet criticized the scale for being too subjective, for having items reflecting formal education, and for using a yes or no format on many questions (DuBois, 1970). But he was much impressed with the idea of using a battery of tests, a feature which he adopted in his 1905 scales.

In 1904, the Minister of Public Instruction in Paris appointed a commission to decide upon the educational measures that should be undertaken with those children who could not profit from regular instruction. The commission concluded that medical and educational examinations should be used to identify those children who could not learn by the ordinary methods. Furthermore, it was determined that these children should be removed from their regular classes and given special instruction suitable to their more limited intellectual prowess. This was the beginning of the special education classroom.

It was evident that a means of selecting children for such special placement was needed, and Binet and his colleague Simon were called upon to develop a practical tool for just this purpose. Thus arose the first formal scale for assessing the intelligence of children.

Goodenough (1949) has outlined the four ways in which the 1905 scale differed from those which had been previously constructed.

1. It made no pretense of measuring precisely any single faculty. Rather, it was aimed at assessing the child's general mental development with a heterogeneous group of tasks. Thus, the aim was not measurement, but classification.
2. It was a brief and practical test. The test took less than an hour to administer and required little in the way of equipment.
3. It measured directly what Binet and Simon regarded as the essential factor of intelligence—practical judgment—rather than wasting time with lower-level abilities involving sensory, motor, and perceptual elements. They took a pragmatic view of intelligence:

> There is in intelligence, it seems to us, a fundamental agency the lack or alteration of which has the greatest importance for practical life; that is judgement, otherwise known as good sense, practical sense, initiative, or the faculty of adapting oneself. To judge well, to understand well, to reason well—these are the essential wellsprings of intelligence. (Binet and Simon, 1905; as translated in Fancher, 1985)

4. The items were arranged by approximate level of difficulty instead of content. A rough standardization had been done with 50 normal children ranging in age from three to 11 years and several subnormal and retarded children as well.

The 30 tests on the 1905 scale ranged from utterly simple sensory tests to quite complex verbal abstractions. Thus, the scale was appropriate for assessing the entire gamut of intelligence—from severe mental retardation to high levels of giftedness. The entire scale is outlined in Table 1.1.

TABLE 1.1 The 1905 Binet-Simon Scale

1. Follows a moving object with the eyes.
2. Grasps a small object which is touched.
3. Grasps a small object which is seen.
4. Recognizes the difference between a square of chocolate and a square of wood.
5. Finds and eats a square of chocolate wrapped in paper.
6. Executes simple commands and imitates simple gestures.
7. Points to familiar named objects, e.g., "Show me the cup."
8. Points to objects represented in pictures, e.g., "Put your finger on the window."
9. Names objects in pictures, e.g., "What is this?" [examiner points to a picture of a sign].
10. Compares two lines of markedly unequal length.
11. Repeats three spoken digits.
12. Compares two weights.
13. Shows susceptibility to suggestion.
14. Defines common words by function.
15. Repeats a sentence of 15 words.
16. Tells how two common objects are different, e.g., "paper and cardboard."
17. Names from memory as many as possible of 13 objects displayed on a board for 30 seconds. [This test was later dropped because it permitted too many possibilities for distraction.]
18. Reproduces from memory two designs shown for 10 seconds.
19. Repeats a longer series of digits than in item 11 to test immediate memory.
20. Tells how two common objects are alike, e.g., "butterfly and flea."
21. Compares two lines of slightly unequal length.
22. Compares five blocks to put them in order of weight.
23. Indicates which of the previous five weights the examiner has removed.
24. Produces rhymes, e.g., "What rhymes with 'school'?"
25. A word completion test based on those proposed by Ebbinghaus.
26. Puts three nouns, e.g., "Paris, river, fortune" (or three verbs) in a sentence.
27. Responds to 25 abstract (comprehension) questions, e.g., "When a person has offended you, and comes to offer his apologies, what should you do?"
28. Reverses the hands of a clock.
29. After paper folding and cutting, draws the form of the resulting holes.
30. Defines abstract words by designating the difference between, e.g., "boredom and weariness."

Source: Based on translations in Jenkins and Paterson (1961) and Jensen (1980).

Except for the very simplest tests, which were designed for the classification of very low-grade *idiots* (an unfortunate diagnostic term that has since been dropped), the tests were heavily weighted toward verbal skills, reflecting Binet's departure from the Galtonian tradition.

An interesting point that is often overlooked by contemporary students of psychology is that Binet and Simon did not offer a precise method for arriving at a total score on their 1905 scale. It is well to remember that their purpose was classification, not measurement, and that their motivation was entirely humanitarian, namely, to identify those children who needed special educational placement. By contemporary standards, it is difficult to accept the fuzziness inherent in such an approach, but that may reflect a modern penchant for quantification more than a weakness in the 1905 scale. In fact, their scale was popular among educators in Paris. And, even with the absence of precise quantification, the approach was successful in selecting candidates for special classes.

THE REVISED SCALES AND THE ADVENT OF IQ

In 1908, Binet and Simon published a revision of the 1905 scale. In the earlier scale, more than half the items had been designed for the very retarded, yet the major diagnostic decisions involved older children and those with borderline intellect. To remedy this imbalance, most of the very simple items were dropped and new items were added at the higher end of the scale. The 1908 scale had 58 problems or tests, almost double the number from 1905. Several new tests were added, many of which are still used today: reconstructing scrambled sentences, copying a diamond, and executing a sequence of three commands. Some of the items were absurdities that the children had to detect and explain. One such item was amusing to French children: "The body of an unfortunate girl was found, cut into 18 pieces. It is thought that she killed herself." However, this item was very upsetting to some American subjects, demonstrating the importance of cultural factors in intelligence (Fancher, 1985).

The major innovation of the 1908 scale was the introduction of the concept of mental level. The tests had been standardized on about 300 normal children between the ages of 3 and 13 years. This allowed Binet and Simon to order the tests according to the age level at which they were typically passed. Whichever items were passed by 80 to 90 percent of the 3-year-olds were placed in the 3-year level, and similarly on up to age 13. Binet and Simon also devised a rough scoring system whereby a basal age was first determined from the age level at which not more than one test was failed. For each five tests that were passed at levels above the basal, a full year of mental level was granted. Insofar as partial years of mental level were not credited and the various age levels had anywhere from three to eight tests, the method left much to be desired.

In 1911, a third revision of the Binet-Simon scales appeared. Each age level now had exactly five tests. The scale was also extended into the adult range. And with some reluctance, Binet introduced new scoring methods that allowed for one-fifth of a year for each subtest passed beyond the basal level. In his writings, Binet emphasized strongly that the child's exact mental level should not be taken too seriously as an absolute measure of intelligence.

Nonetheless, the idea of deriving a mental level was a monumental development that was to influence the character of intelligence testing throughout the twentieth century. Within months, what Binet called mental level was being translated as mental age. And testers everywhere, including Binet himself, were comparing a child's mental age with the child's chronological age. Thus, a 9-year-old who was functioning at the mental level (or mental age) of a 6-year-old was retarded by three years. Very shortly, Stern (1912) pointed out that being retarded by three years had different meanings at different ages. A 5-year-old functioning at the 2-year-old level was more impaired than a 13-year-old functioning at the 10-year-old level. Stern suggested that an intelligence quotient computed from the mental age divided by the chronological age would give a better measure of the relative functioning of a subject compared to his or her same-aged peers.

In 1916, Terman and his associates at Stanford revised the Binet-Simon scales, producing the Stanford-Binet, a successful test that is discussed in a later chapter. Terman suggested multiplying the intelligence quotient by 100 to remove fractions; he was also the first person to use the abbreviation *IQ*. Thus was born one of the most popular and controversial concepts in the history of psychology. Binet died in 1911 before the IQ swept American testing, so we will never know what he would have thought of this new development based on his scales. However, Simon, his collaborator, later called the concept of IQ a "betrayal" of their scale's original objectives (Fancher, 1985, p. 104), and we can assume from Binet's humanistic concern that he might have held a similar opinion.

SUMMARY

1. For better or for worse, psychological test results possess the power to alter lives. A review of historical trends is crucial if we desire to comprehend the contemporary influence of psychological tests.

2. Rudimentary forms of testing date back to 2200 B.C. in China. The Chinese emperors used grueling written exams to select officials for civil service.

3. In the mid- to late 1800s, several physicians and psychiatrists developed standardized procedures to reveal the nature and extent of symptoms in the mentally ill and brain-injured. For example, in 1885, Hubert von Grashey developed the precursor to the memory drum to test the visual recognition skill of brain-injured patients.

4. Modern psychological testing owes its inception to the era of brass instruments psychology that flourished in Europe during the late 1800s. By testing sensory thresholds and reaction times, pioneer test developers such as Sir Francis Galton demonstrated that it was possible to measure the mind in an objective and replicable manner.

5. Wilhelm Wundt founded the first psychological laboratory in 1879 in Leipzig, Germany. Included among his earlier investigations was his 1862 attempt to measure the speed of thought with the thought meter, a calibrated pendulum with needles sticking off from each side.

6. The first reference to mental tests occurred in 1890 in a classic paper by James McKeen Cattell, an American psychologist who had studied with Galton. Cattell imported the brass instruments approach to the United States.

7. One of Cattell's students, Clark Wissler, showed that reaction time and sensory discrimination measures did not correlate with college grades, thereby redirecting the mental-testing movement away from brass instruments.

8. In the late 1800s, a newfound humanism toward the mentally retarded, reflected in the diagnostic and remedial work of French physicians Esquirol and Seguin, helped create the necessity for early intelligence tests.

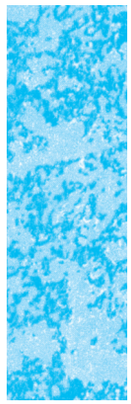
9. Alfred Binet, who was to invent the first true intelligence test, began his career by studying hysterical paralysis with the French neurologist Charcot. Binet's claim that magnetism could cure hysteria was, to his pained embarrassment, disproved. Shortly thereafter, he switched interests and conducted sensory-perceptual studies, using his children as subjects.

10. In 1905, Binet and Simon developed the first useful intelligence test in Paris, France. Their simple 30-item measure of mainly higher mental functions helped identify schoolchildren who could not profit from regular instruction. Curiously, there was no method for scoring the test.

11. In 1908, Binet and Simon published a revised 58-item scale that incorporated the concept of mental level. In 1911, a third revision of the Binet-Simon scales appeared. Each age level now had exactly five tests; the scale extended into the adult range.

12. In 1912, Stern proposed dividing the mental age by the chronological age to obtain an intelligence quotient. In 1916, Terman suggested multiplying the intelligence quotient by 100 to remove fractions. Thus was born the concept of IQ.

TOPIC 1B Early Testing in the United States



The Binet-Simon scales helped solve a practical social quandary, namely, how to identify children who needed special schooling. With this successful application of a mental test, psychologists realized that their inventions could have pragmatic significance for many different segments of society. Almost immediately, psychologists in the United States adopted a utilitarian focus. Intelligence testing was embraced by many as a reliable and objective response to perceived social problems such as the identification of immigrants with mental retardation and the quick, accurate classification of Army recruits (Boake, 2002).

Whether these early tests really solved social dilemmas—or merely exacerbated them—is a fiercely debated issue reviewed in the following sections. One thing is certain: The profusion of tests developed early in the twentieth century helped shape the character of contemporary tests. A review of these historical trends will aid in the comprehension of the nature of modern tests and a better appreciation of the social issues raised by them.

EARLY USES AND ABUSES OF TESTS IN THE UNITED STATES

First Translation of the Binet-Simon Scale

In 1906, Henry H. Goddard was hired by the Vineland Training School in New Jersey to do research on the classification and education of "feebleminded" children. He soon realized that a diagnostic instrument would be required and was therefore pleased to read of the 1908 Binet-Simon scale. He quickly set about translating the scale, making minor changes so that it would be applicable to American children (Goddard, 1910a).

Goddard (1910b) tested 378 residents of the Vineland facility and categorized them by diagnosis and mental age. He classified 73 residents as *idiots* because their mental age was 2 years or lower; 205 residents were termed *imbeciles* with mental age of 3 to 7 years; and 100 residents were deemed *feebleminded* with mental age of 8 to 12 years. It is instructive to note that originally neutral and descriptive terms for portraying levels of mental retardation—idiot, imbecile, and feebleminded—

have made their way into the everyday lexicon of pejorative labels. In fact, Goddard made his own contribution by coining the diagnostic term *moron* (from the Greek *moronia,* meaning "foolish").

Goddard (1911) also tested 1,547 normal children with his translation of the Binet-Simon scales. He considered children whose mental age was four or more years behind their chronological age to be feebleminded—these constituted 3 percent of his sample. Considering that all of these children were found outside of institutions for the retarded, 3 percent is rather an alarming rate of mental deficiency. Goddard (1911) was of the opinion that these children should be segregated so that they would be prevented from "contaminating society." These early studies piqued Goddard's curiosity about "feebleminded" citizenry and the societal burdens they imposed. He also gained a reputation as one of the leading experts on the use of intelligence tests to identify persons with impaired intellect. His talents were soon in heavy demand.

The Binet-Simon and Immigration

In 1910, Goddard was invited to Ellis Island by the commissioner of immigration to help make the examination of immigrants more accurate. A dark and foreboding folklore had grown up around mental deficiency and immigration in the early 1900s:

> It was believed that the feebleminded were degenerate beings responsible for many if not most social problems; that they reproduced at an alarming rate and menaced the nation's overall biological fitness; and that their numbers were being incremented by undesirable "new" immigrants from southern and eastern European countries who had largely supplanted the "old" immigrants from northern and western Europe. (Gelb, 1986)

Initially, Goddard was unconcerned about the supposed threat of feeblemindedness posed by the immigrants. He wrote that adequate statistics did not exist and that the prevalent opinions about undue percentages of mentally defective immigrants were "grossly overestimated" (Goddard, 1912). However, with repeated visits to Ellis Island, Goddard became convinced that the rates of feeblemindedness were much higher than estimated by the physicians who staffed the immigration service. Within a year, he reversed his opinions entirely and called for congressional funding so that Ellis Island could be staffed with experts trained in the use of intelligence tests. In the following decade, Goddard became an apostle for the use of intelligence tests to identify feebleminded immigrants. Although he wrote that the rates of mentally deficient immigrants were "alarming," he did not join the popular call for immigration restriction (Gelb, 1986).

The story of Goddard and his concern for the "menace of feeblemindedness," as Gould (1981) has satirically put it, is often ignored or downplayed in books on psychological testing. The majority of textbooks on testing do not mention or refer to Goddard at all. The few books that do mention him usually state that Goddard "used the tests in institutions for the retarded," which is surely an understatement. In his influential *History of Psychological Testing,* DuBois (1970) has a portrait of Goddard but devotes less than one line of text to him.

The fact is that Goddard was one of the most influential American psychologists of the early 1900s. Any thoughtful person must therefore wonder why so many contemporary authors have ignored or slighted the person who first translated and applied Binet's tests in the United States. We will attempt an answer here, based in part on Goddard's original writing, but also relying upon Gould's (1981) critique of Goddard's voluminous writings on mental deficiency and intelligence testing. We refer to Gelb's (1986) more sympathetic portrayal of Goddard as well.

Perhaps Goddard has been ignored in the textbooks because he was a strict hereditarian who conceived of intelligence in simple-minded Mendelian terms. No doubt his call for colonization of "morons" so as to restrict their breeding has won him contemporary disfavor as well. And his insistence that much undesirable behavior—crime, alcoholism, prostitution—was due to inherited mental deficiency also does not sit well with the modern environmentalist position.

However, the most likely reason that modern authors have ignored Goddard is that he exempli-

fied a large number of early, prominent psychologists who engaged in the blatant misuse of intelligence testing. In his efforts to demonstrate that high rates of immigrants with mental retardation were entering the United States each day, Goddard sent his assistants to Ellis Island to administer his English translation of the Binet-Simon tests to newly arrived immigrants. The tests were administered through a translator, not long after the immigrants walked ashore. We can guess that many of the immigrants were frightened, confused, and disoriented. Thus, a test devised in French, then translated to English was, in turn, retranslated back to Yiddish, Hungarian, Italian, or Russian; administered to bewildered farmers and laborers who had just endured an Atlantic crossing; and interpreted according to the original French norms.

What did Goddard find and what did he make of his results? In small samples of immigrants (22 to 50), his assistants found 83 percent of the Jews, 80 percent of the Hungarians, 79 percent of the Italians, and 87 percent of the Russians to be feebleminded, that is, below age 12 on the Binet-Simon scales (Goddard, 1917). His interpretation of these findings is, by turns, skeptically cautious and then provocatively alarmist. In one place he claims that his study "makes no determination of the actual percentage, even of these groups, who are feebleminded." Yet, later in the report he states that his figures would only need to be revised by "a relatively small amount" in order to find the actual percentages of feeblemindedness among immigrant groups. Further, he concludes that the intelligence of the average immigrant is low, "perhaps of moron grade," but then goes on to cite environmental deprivation as the primary culprit. Simultaneously, Goddard appears to favor deportation for low IQ immigrants but also provides the humanitarian perspective that we might be able to use "moron laborers" if only "we are wise enough to train them properly."

There is much, much more to the Goddard era of early intelligence testing, and the interested reader is urged to consult Gould (1981) and Gelb (1986). The most important point that we wish to stress here is that—like many other early psychologists—Goddard's scholarly views were influenced by the social ideologies of his time. Finally, Goddard was a complex scholar who refined and contradicted his professional opinions on numerous occasions. One ironic example: After the damage was done and his writings had helped restrict immigration, Goddard (1928) recanted, concluding that feeblemindedness was not incurable, and that the feebleminded did not need to be segregated in institutions.

The Goddard chapter in the history of testing serves as a reminder that even well-meaning persons operating within generally accepted social norms can misuse psychological tests. We need be ever mindful that disinterested "science" can be harnessed to the goals of a pernicious social ideology.

THE INVENTION OF NONVERBAL TESTS IN THE EARLY 1900S

Because of the heavy emphasis of the Binet-Simon scales upon verbal skills, many psychologists realized that this new measuring device was not entirely appropriate for non-English-speaking subjects, illiterates, and those with speech and hearing impairments. A spate of performance scales therefore arose in the decade following Goddard's 1908 translation of the Binet-Simon. Only a brief chronology of nonverbal tests will be supplied here. The interested reader should consult DuBois (1970). In this listing of early performance tests, the reader will surely recognize many instruments and subtests that are still used today.

The earliest of the performance measures was the Seguin form board, an upright stand with depressions into which ten blocks of varying shapes could be fitted. This had been used by Seguin as a training device for individuals with mental retardation, but was subsequently developed as a test by Goddard, and then standardized by R. H. Sylvester (1913). This identical board is still used, with the subject blindfolded, in the Halstead-Reitan neuropsychological test battery (Reitan & Wolfson, 1985).

Knox (1914) devised several performance tests for use with Ellis Island immigrants. His tests

required absolutely no verbal responses from subjects. The examiner demonstrated each task nonverbally to ensure that the subjects understood the instructions. Included in his tests were a simple wooden puzzle (which Knox referred to as the "moron" test) and the same digit-symbol substitution test which is now found on most of the Wechsler scales of intelligence.

Several other early performance tests are worthy of brief mention because they have survived to the present day in revised form. Pintner and Paterson (1917) invented a 15-part scale of performance tests that used several form boards, puzzles, and object assembly tests. The object assembly test—reassembling cut-up cardboard versions of common objects such as a horse—is a mainstay of several contemporary intelligence tests. The Kohs Block Design test (Kohs, 1920), which required the subject to assemble painted blocks to resemble a pattern, is well known to any modern tester who uses the Wechsler scales. The Porteus Maze Test (Porteus, 1915) is a graded series of mazes for which the subject must avoid dead ends while tracing a path from beginning to end. This is a fine instrument that is still available today, but underused.

THE STANFORD-BINET: THE EARLY MAINSTAY OF IQ

While it was Goddard who first translated the Binet scales in the United States, it was Stanford professor Lewis M. Terman (1857–1956) who popularized IQ testing with his revision of the Binet scales in 1916. The new Stanford-Binet, as it was called, was a substantial revision, not just an extension, of the earlier Binet scales. Among the many changes that led to the unquestioned prestige of the Stanford-Binet was the use of the now familiar IQ for expressing test results. The number of items was increased to 90, and the new scale was suitable for those with mental retardation, children, and both normal and "superior" adults. In addition, the Stanford-Binet had clear and well-organized instructions for administration and scoring. Great care had been taken in securing a representative sample of subjects for use in the standardization of the test. As Goodenough (1949) notes: "The publication of the Stanford Revision marked the end of the initial period of experimentation and uncertainty. Once and for all, intelligence testing had been put on a firm basis."

The Stanford-Binet was the standard of intelligence testing for decades. New tests were always validated in terms of their correlations with this measure. It continued its preeminence through revisions in 1937, and 1960, by which time the Wechsler scales (Wechsler, 1949, 1955) had begun to compete with it. The latest revision of the Stanford-Binet was completed in 2003. This test and the Wechsler scales are discussed in detail in a later chapter. It is worth mentioning here that the Wechsler scales became a quite popular alternative to the Stanford-Binet mainly because they provided more than just an IQ score. In addition to Full Scale IQ, the Wechsler scales provided ten to twelve subtest scores, and a Verbal and Performance IQ. By contrast, the earlier versions of the Stanford-Binet supplied only a single overall summary score, the global IQ.

GROUP TESTS AND THE CLASSIFICATION OF WWI ARMY RECRUITS

Given the American penchant for efficiency, it was only natural that researchers would seek group mental tests to supplement the relatively time-consuming individual intelligence tests imported from France. Among the first to develop group tests was Pyle (1913), who published schoolchildren norms for a battery consisting of such well-worn measures as memory span, digit-symbol substitution, and oral word association (quickly writing down words in response to a stimulus word). Pintner (1917) revised and expanded Pyle's battery, adding to it a timed cancellation test in which the child crossed out the letter *a* wherever it appeared in a body of text.

But group tests were slow to catch on, partly because the early versions still had to be scored

laboriously by hand. The idea of a completely objective test with a simple scoring key was inconsistent with tests such as logical memory for which the judgment of the examiner was required in scoring. Most amazing of all—at least to anyone who has spent any time as a student in American schools—the multiple-choice question was not yet in general use.

The slow pace of developments in group testing picked up dramatically as the United States entered World War I in 1917. It was then that Robert M. Yerkes, a well-known psychology professor at Harvard, convinced the U.S. government and the Army that all of its 1.75 million recruits should be given intelligence tests for purposes of classification and assignment (Yerkes, 1919). Immediately upon being commissioned into the Army as a colonel, Yerkes assembled a Committee on the Examination of Recruits, which met at the Vineland school in New Jersey to develop the new group tests for the assessment of Army recruits. Yerkes chaired the committee; other famous members included Goddard and Terman.

Two group tests emerged from this collaboration: the Army Alpha and the Army Beta. It would be difficult to overestimate the influence of the Alpha and Beta upon subsequent intelligence tests. The format and content of these tests inspired developments in group and individual testing for decades to come. We discuss these tests in some detail so that the reader can appreciate their influence on modern intelligence tests.

The Army Alpha and Beta Examinations

The Alpha was based on the then unpublished work of Otis (1918) and consisted of eight verbally loaded tests for average and high-functioning recruits. The eight tests were: (1) following oral directions, (2) arithmetical reasoning, (3) practical judgment, (4) synonym–antonym pairs, (5) disarranged sentences, (6) number series completion, (7) analogies, and (8) information. Figure 1.1 lists some typical items from the Army Alpha examination.

The Army Beta was a nonverbal group test designed for use with illiterates and recruits whose first language was not English. It consisted of various visual-perceptual and motor tests such as tracing a path through mazes and visualizing the correct number of blocks depicted in a three-dimensional drawing. Figure 1.2 depicts the blackboard demonstrations for all eight parts of the Beta examination.

In order to accommodate illiterate subjects and recent immigrants who did not comprehend English, Yerkes instructed the examiners to use largely pictorial and gestural methods for explaining the tests to the prospective Army recruits. The examiner and an assistant stood atop a platform at the front of the class and engaged in pantomime to explain each of the eight tests. We reproduce here the exact instructions for one test so that the reader can appraise the likely effects of the testing procedures upon Beta results. Keep in mind that many recruits could not see or hear the examiner well, and that some had never taken a test before. Here is how the examiners introduced test 6, picture completion, to each new roomful of potential recruits:

> "This is test 6 here. Look. A lot of pictures." After everyone has found the place, "Now watch." Examiner points to hand and says to demonstrator, "Fix it." Demonstrator does nothing, but looks puzzled. Examiner points to the picture of the hand, and then to the place where the finger is missing and says to demonstrator, "Fix it; fix it." Demonstrator then draws in finger. Examiner says "That's right." Examiner then points to fish and place for eye and says, "Fix it." After demonstrator has drawn missing eye, examiner points to each of the four remaining drawings and says, "Fix them all." Demonstrator works samples out slowly and with apparent effort. When the samples are finished examiner says, "All right. Go head. Hurry up!" During the course of this test the orderlies walk around the room and locate individuals who are doing nothing, point to their pages and say, "Fix it. Fix them," trying to set everyone working. At the end of 3 minutes examiner says, "Stop! But don't turn over the page." (Yerkes, 1921)

The Army testing was intended to help segregate and eliminate the mentally incompetent, to classify men according to their mental ability, and to assist in the placement of competent men in

FOLLOWING ORAL DIRECTIONS

Mark a cross in the first and also the third circle:

○ ○ ○ ○ ○

ARITHMETICAL REASONING

Solve each problem:

How many men are 5 men and 10 men? Answer ( )

If 3 1/2 tons of coal cost $21, what will 5 1/2 tons cost? Answer ( )

PRACTICAL JUDGMENT

Why are high mountains covered with snow? Because

- ☐ they are near the clouds
- ☐ the sun shines seldom on them
- ☐ the air is cold there

SYNONYM–ANTONYM PAIRS

Are these words the same or opposite?

largess—donation same? or opposite?

accumulate—dissipate same? or opposite?

DISARRANGED SENTENCES

Can these words be rearranged to form a sentence?

envy bad malice traits are and true? or false?

NUMBER SERIES COMPLETION

Complete the series: 3 6 8 16 18 36 . . . . . .

ANALOGIES

Which choice completes the analogy?

tears—sorrow :: laughter— joy smile girls grin

granary—wheat :: library— desk books paper librarian

INFORMATION

Choose the best alternative:

The pancreas is in the abdomen head shoulder neck

The Battle of Gettysburg was fought in 1863 1813 1778 1812

Note: Examinees received verbal instructions for each subtest.

FIGURE 1.1 Sample Items from the Army Alpha Examination
Source: Reprinted from Yerkes, R. M. (Ed.). (1921). *Psychological examining in the United States Army. Memoirs of the National Academy of Sciences, Volume 15.* With permission from the National Academy of Sciences, Washington, DC.

responsible positions (Yerkes, 1921). However, it is not really clear whether the Army made much use of the masses of data supplied by Yerkes and his eager assistants. A careful reading of his memoirs reveals that Yerkes did little more than produce favorable testimonials from high-ranking officers. In the main, his memoirs say that the Army could have saved millions of dollars and increased its efficiency, if the testing data had been used.

To some extent, the mountains of test data had little practical impact on the efficiency of the Army

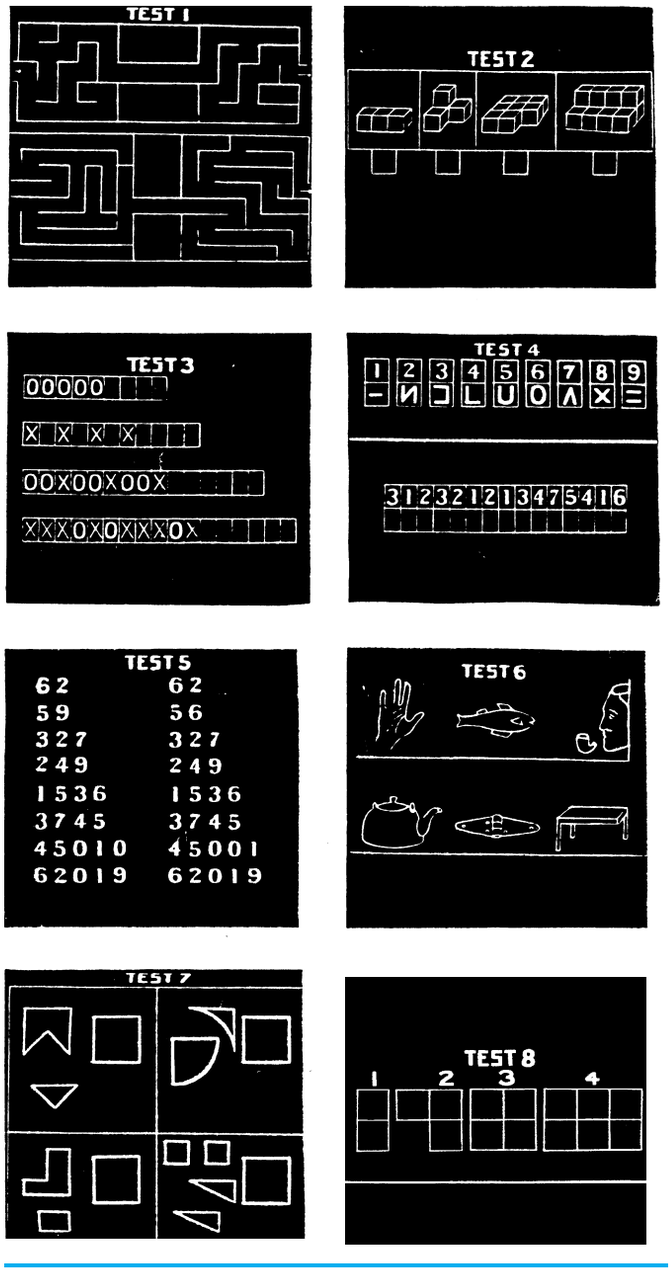


FIGURE 1.2
The Blackboard Demonstrations for All Eight Parts of the Beta Examination
Source: Reprinted from Yerkes, R. M. (Ed.). (1921). *Psychological examining in the United States Army. Memoirs of the National Academy of Sciences, Volume 15.* With permission from the National Academy of Sciences, Washington, DC.

because of the resistance of the military mind to scientific innovation. However, it is also true that the Army brass had good reason to doubt the validity of the test results. For example, an internal memorandum described the use of pantomime in the instructions to the nonverbal Beta examination:

> For the sake of making results from the various camps comparable, the examiners were ordered to follow a certain detailed and specific series of ballet antics, which had not only the merit of being perfectly incomprehensible and unrelated to mental testing, but also lent a highly confusing and distracting mystical atmosphere to the whole performance, effectually preventing all approach to the attitude in which a subject should be while having his soul tested. (cited in Samelson, 1977)

In addition, the testing conditions left much to be desired, with wave upon wave of recruits ushered in one door, tested, and virtually shoved out the other side. Tens of thousands of recruits received a literal zero for many subtests, not because they were retarded but because they couldn't fathom the instructions to these enigmatic new instruments. Many recruits fell asleep while the testers gave esoteric and mysterious pantomime instructions.

On the positive side, the Army testing provided psychologists with a tremendous amount of experience in the psychometrics of test construction. Thousands of correlation coefficients were computed, including the prominent use of multiple correlations in the analysis of test data. Test construction graduated from an art to a science in a few short years.

The Army Tests and Ethnic Differences

Unfortunately, the Army test results were sometimes used to substantiate prejudices about various racial and ethnic groups rather than to dispassionately investigate the causes of group differences. For example, in his influential book *A Study of American Intelligence,* Brigham (1923) undertook a massive analysis of Alpha and Beta scores for Nordic, Mediterranean, and Alpine immigrants. The text is stuffed with ostensibly objective tables and charts comparing racial and ethnic groups. For example, one curious figure in his book depicts the proportion of each immigration sample at or below the average of the African American draft. Brigham concluded that African Americans, Mediterranean immigrants, and Alpine immigrants were intellectually inferior. He sounded a dire warning that racial intermixture would inevitably cause a deterioration of American intelligence. For example, the caption to one graph reads, in part:

> The distributions of the intelligence scores of the entire Nordic group, the combined Mediterranean and Alpine groups, and the negro draft. The process of racial intermixture cannot result in anything but an average of these elements, with the resulting deterioration of American intelligence. (Brigham, 1923)

Seven years later, Brigham (1930) forthrightly disavowed his earlier views. He cited cultural and language differences as the likely cause of ethnic and racial disparities on the Army tests. He asserted that comparative studies of national and racial groups could not be made with existing tests and concluded that his earlier findings were "without foundation" (Brigham, 1930).

EARLY EDUCATIONAL TESTING

For good or for ill, Yerkes's grand scheme for testing Army recruits helped to usher in the era of group tests. After WWI, inquiries rushed in from industry, public schools, and colleges about the potential applications of these straightforward tests that almost anyone could administer and score (Yerkes, 1921). The psychologists who had worked with Yerkes soon left the service and carried with them to industry and education their newfound notion of paper-and-pencil tests of intelligence.

The Army Alpha and Beta were also released for general use. These tests quickly became the prototypes for a large family of group tests and influenced the character of intelligence tests, college entrance examinations, scholastic achievement tests, and aptitude tests. To cite just one specific consequence of the Army testing, the National Research Council, a government organization of scientists, devised the National Intelligence Test, which was eventually given to 7 million children in the United States during the 1920s. Thus, such well-known tests as the Wechsler scales, the Scholastic Aptitude Tests, and the Graduate Record Exam actually have roots that reach back to Yerkes,

Otis, and the mass testing of Army recruits during WWI.

The College Entrance Examination Board (CEEB) was established at the turn of the twentieth century to help avoid duplication in the testing of applicants to U.S. colleges. The early exams had been of the short answer essay format, but this was to change quickly when C. C. Brigham, a disciple of Yerkes, became CEEB secretary after WWI. In 1925, the College Board decided to construct a scholastic aptitude test for use in college admissions (Goslin, 1963). The new tests reflected the now familiar objective format of unscrambling sentences, completing analogies, and filling in the next number in a sequence. Machine scoring was introduced in the 1930s, making objective group tests even more efficient than before. These tests then evolved into the present College Board tests, in particular, the Scholastic Aptitude Tests, now known as the Scholastic Assessment Tests.

The functions of the CEEB were later subsumed under the nonprofit Educational Testing Service (ETS). The ETS directed the development, standardization, and validation of such well-known tests as the Graduate Record Examination, the Law School Admissions Test, and the Peace Corps Entrance Tests.

Meanwhile, Terman and his associates at Stanford were busy developing standardized achievement tests. The Stanford Achievement Test (SAchT) was first published in 1923; a modern version of it is still in wide use today. From the very beginning, the SAchT incorporated such modern psychometric principles as norming the subtests so that within-subject variability could be assessed and selecting a very large and representative standardization sample.

THE DEVELOPMENT OF APTITUDE TESTS

Aptitude tests measure more specific and delimited abilities than intelligence tests. Traditionally, intelligence tests assess a more global construct such as general intelligence, although there are exceptions to this trend that will be discussed later. By contrast, a single aptitude test will measure just one ability domain, and a multiple aptitude test battery will provide scores in several distinctive ability areas.

The development of aptitude tests lagged behind that of intelligence tests for two reasons, one statistical, the other social. The statistical problem was that a new technique, factor analysis, was often needed to discern which aptitudes were primary and therefore distinct from each other. Research on this question had been started quite early by Spearman (1904) but was not refined until the 1930s (Spearman, 1927; Kelley, 1928; Thurstone, 1938). This new family of techniques, factor analysis, allowed Thurstone to conclude that there were specific factors of primary mental ability such as verbal comprehension, word fluency, number facility, spatial ability, associative memory, perceptual speed, and general reasoning (Thurstone, 1938; Thurstone & Thurstone, 1941). More will be said about this in the later chapters on intelligence and ability testing. The important point here is that Thurstone and his followers thought that global measures of intelligence did not, so to speak, "cut nature at its joints." As a result, it was felt that such measures as the Stanford-Binet were not as useful as multiple aptitude test batteries in determining a person's intellectual strengths and weaknesses.

The second reason for the slow growth of aptitude batteries was the absence of a practical application for such refined instruments. It was not until WWII that a pressing need arose to select candidates who were highly qualified for very difficult and specialized tasks. The job requirements of pilots, flight engineers, and navigators were very specific and demanding. A general estimate of intellectual ability, such as provided by the group intelligence tests used in WWI, was not sufficient to choose good candidates for flight school. The armed forces solved this problem by developing a specialized aptitude battery of 20 tests that was administered to men who passed preliminary screening tests. These measures proved invaluable in selecting pilots, navigators, and bombadiers, as reflected in the much lower washout rates of men

selected by test battery instead of the old methods (Goslin, 1963). Such tests are still used widely in the armed services.

PERSONALITY AND VOCATIONAL TESTING AFTER WWI

While such rudimentary assessment methods as the free association technique had been used before the turn of the twentieth century by Galton, Kraepelin, and others, it was not until WWI that personality tests emerged in a form resembling their contemporary appearance. As has happened so often in the history of testing, it was once again a practical need that served as the impetus for this new development. Modern personality testing began when Woodworth attempted to develop an instrument for detecting Army recruits who were susceptible to psychoneurosis. Virtually all the modern personality inventories, schedules, and questionnaires owe a debt to Woodworth's Personal Data Sheet (1919).

The Personal Data Sheet consisted of 116 questions that the subject was to answer by underlining *Yes* or *No*. The questions were exclusively of the "face obvious" variety and, for the most part, involved fairly serious symptomatology. Representative items included:

- Do ideas run through your head so that you cannot sleep?
- Were you considered a bad boy?
- Are you bothered by a feeling that things are not real?
- Do you have a strong desire to commit suicide?

Readers familiar with the Minnesota Multiphasic Personality Inventory (MMPI) must surely recognize the debt that this more recent inventory has to Woodworth's instrument.

From his account of how the Personal Data Sheet was developed (Woodworth, 1951), it is clear that Woodworth took great care in the selection of items. In other respects, though, this instrument embodies a large dose of psychometric credulity. The most serious problem is simply that a disturbed subject motivated to look good could do so without detection; likewise, a normal subject with a *fake bad* mentality might be categorized as unfit for service. Modern instruments such as the MMPI have incorporated various validity scales for detecting such response tendencies. The Personal Data Sheet, by contrast, was predicated on the assumption that subjects would be honest when responding to the questions.

The next major development was an inventory of neurosis, the Thurstone Personality Schedule (Thurstone & Thurstone, 1930). After first culling hundreds of items answerable in the yes-no-? manner from Woodworth's inventory and other sources, Thurstone rationally keyed items in terms of how the neurotic would typically answer them. Reflecting Thurstone's penchant for statistical finesse, this inventory was one of the first to use the method of internal consistency whereby each prospective item was correlated with the total score on the tentatively identified scale to determine whether it belonged on the scale.

From the Thurstone test sprang the Bernreuter Personality Inventory (Bernreuter, 1931). It was a little more refined than its Thurstone predecessor, measuring four personality dimensions: neurotic tendency, self-sufficiency, introversion-extroversion, and dominance-submission. A major innovation in test construction was that a single test item could contribute to more than one scale.

The Allport-Vernon Study of Values was also published in 1931 (Allport & Vernon, 1931). This test was quite different from the others in that it measured values instead of psychopathology. Furthermore, it adopted a new scoring method, the ipsative approach, in which the respondent was compared only with himself or herself regarding the balance of importance given to six basic values: theoretical, economic, aesthetic, social, political, and religious. The test was devised in such a manner that subjects were required to make choices between the six values in specific situations. As a consequence, the average on the six scales was always the same for each subject. A weakness in one value was compensated for by a strength in some other value. Thus, only the relative peaks and valleys were of interest.

Any chronology of self-report inventories must surely include the Minnesota Multiphasic Personality Inventory, or MMPI (Hathaway & McKinley, 1940). This test and its revision, the MMPI-2, are discussed in detail later. It will suffice for now to point out that the scales of the MMPI were constructed by the method that Woodworth pioneered, contrasting the responses of normal and psychiatrically disturbed subjects. In addition, the MMPI introduced the use of validity scales to determine fake bad, fake good, and random response patterns.

THE ORIGINS OF PROJECTIVE TESTING

The projective approach originated with the word association method pioneered by Francis Galton in the late 1800s. Galton gave himself four seconds to come up with as many associations as possible to a stimulus word, and then categorized his associations as parrotlike, image-mediated, or histrionic representations. This latter category convinced him that mental operations "sunk wholly below the level of consciousness" were at play. Some historians have even speculated that Freud's application of free association as a therapeutic tool in psychoanalysis sprang from Galton's paper published in *Brain* in 1879 (Forrest, 1974).

Galton's work was continued in Germany by Wundt and Kraepelin, and finally brought to fruition by Jung (1910). Jung's test consisted of 100 stimulus words. For each word, the subject was to reply as quickly as possible with the first word coming to mind. Kent and Rosanoff (1910) gave the association method a distinctively American flavor by tabulating the reactions of 1,000 normal subjects to a list of 100 stimulus words. These tables were designed to provide a basis for comparing the reactions of normal and "insane" subjects.

While the Americans were pursuing the empirical approach to objective personality testing, a young Swiss psychiatrist, Hermann Rorschach (1884–1922), was developing a completely different vehicle for studying personality. Rorschach was strongly influenced by Jungian and psychoanalytic thinking, so it was natural that his new approach focused on the tendency of patients to reveal their innermost conflicts unconsciously when responding to ambiguous stimuli. The Rorschach and other projective tests discussed subsequently were predicated upon the projective hypothesis: When responding to ambiguous or unstructured stimuli, we inadvertently disclose our innermost needs, fantasies, and conflicts.

Rorschach was convinced that people revealed important personality dimensions in their responses to inkblots. He spent years developing just the right set of ten inkblots and systematically analyzed the responses of personal friends and different patient groups (Rorschach, 1921). Unfortunately, he died only a year after his monograph was published, and it was up to others to complete his work. Developments in the Rorschach are reviewed later in the text.

While Rorschach's test was originally developed to reveal the innermost workings of the abnormal subject, the TAT, or Thematic Apperception Test (Morgan & Murray, 1935), was developed as an instrument to study normal personality. Of course, both have since been expanded for testing with the entire continuum of human behavior.

The TAT consists of a series of pictures that largely depict one or more persons engaged in an ambiguous interaction. The subject is shown one picture at a time and told to make up a story about it. He or she is instructed to be as dramatic as possible, to discuss thoughts and feelings, and to describe the past, present, and future of what is depicted in the picture.

Murray (1938) believed that underlying personality needs, such as the need for achievement, would be revealed by the contents of the stories. Although numerous scoring systems were developed, clinicians in the main have relied upon an impressionistic analysis to make sense of TAT protocols. Modern applications of the TAT are discussed in a later chapter.

The sentence completion technique was also begun during this era with the work of Payne (1928). There have been numerous extensions and

variations on the technique, which consists of giving subjects a stem such as "I am bored when _____," and asking them to complete the sentence. Some modern applications are discussed later, but it can be mentioned now that the problem of scoring and interpretation, which vexed early sentence completion test developers, is still with us today.

An entirely new approach to projective testing was taken by Goodenough (1926), who tried to determine not just intellectual level, but also the interests and personality traits of children by analyzing their drawings. Buck's (1948) test, the House-Tree-Person, was a little more standardized and structured and required the subject to draw a house, a tree, and a person. Machover's (1949) *Personality Projection in the Drawing of the Human Figure* was the logical extension of the earlier work. Figure drawing as a projective approach to understanding personality is still used today, and a later chapter discusses modern developments in this practice.

Meanwhile, projective testing in Europe was dominated by the Szondi Test, a wacky instrument based on wholly faulty premises. Lipot Szondi was a Hungarian-born Swiss psychiatrist who believed that major psychiatric disorders were caused by recessive genes. His test consisted of 48 photographs of psychiatric patients divided into six sets of the following eight types: homosexual, epileptic, sadistic, hysteric, catatonic, paranoiac, manic, and depressive (Deri, 1949). From each set of eight pictures, the subject was instructed to select the two pictures he or she liked best and the two disliked most. A person who consistently preferred one kind of picture in the six sets was presumed to have some recessive genes that made him or her have sympathy for the pictured person. Thus, projective preferences were presumed to reveal recessive genes predisposing the individual to specific psychiatric disturbances.

Deri (1949) imported the test to the United States and changed the rationale. She did not argue for a recessive genetic explanation of picture choice but explained such preferences on the basis of unconscious identification with the characteristics of the photographed patients. This was a more palatable theoretical basis for the test than the dubious genetic theories of Szondi. Nonetheless, empirical research cast doubt on the validity of the Szondi Test, and it shortly faded into oblivion (Borstelmann & Klopfer, 1953).

THE DEVELOPMENT OF INTEREST INVENTORIES

While the clinicians were developing measures for analyzing personality and unconscious conflicts, other psychologists were devising measures for guidance and counseling of the masses of more normal persons. Chief among such measures was the interest inventory, which has roots going back to Thorndike's (1912) study of developmental trends in the interests of 100 college students. In 1919–1920, Yoakum developed a pool of 1,000 items relating to interests from childhood through early maturity (DuBois, 1970). Many of these items were incorporated in the Carnegie Interest Inventory. Cowdery (1926–27) improved and refined previous work on the Carnegie instrument by increasing the number of items, comparing responses of three criterion groups (doctors, engineers, and lawyers) with control groups of nonprofessionals, and developing a weighting formula for items. He was also the first psychometrician to realize the importance of cross validation. He tested his new scales on additional groups of doctors, engineers, and lawyers to ensure that the discriminations found in the original studies were reliable group differences rather than capitalizations on error variance.

Edward K. Strong (1884–1963) revised Cowdery's test and devoted 36 years to the development of empirical keys for the modified instrument known as the Strong Vocational Interest Blank (SVIB). Persons taking the test could be scored on separate keys for several dozen occupations, providing a series of scores of immeasurable value in vocational guidance. The SVIB became one of the most widely used tests of all time (Strong, 1927). Its modern version, the Strong In-

terest Inventory, is still widely used by guidance counselors.

For decades the only serious competitor to the SVIB was the Kuder Preference Record (Kuder, 1934). The Kuder differed from the Strong by forcing choices within triads of items. The Kuder was an ipsative test; that is, it compared the relative strength of interests within the individual, rather than comparing his or her responses to various professional groups. More recent revisions of the Kuder Preference Record include the Kuder General Interest Survey and the Kuder Occupational Interest Survey (Kuder, 1966; Kuder & Diamond, 1979; Zytowski, 1985).

SUMMARY OF MAJOR LANDMARKS IN THE HISTORY OF TESTING

We conclude our historical survey of psychological testing with a brief tabular summary of landmark events up to 1950 (Table 1.2). The interested reader can find a more detailed listing—including a chronology of post-1950 developments—in Appendix A.

TABLE 1.2 A Summary of Early Landmarks in the History of Testing

2200 B.C.	Chinese begin civil service examinations.
A.D.1862	Wilhelm Wundt uses a calibrated pendulum to measure the "speed of thought."
1884	Francis Galton administers the first test battery to thousands of citizens at the International Health Exhibit.
1890	James McKeen Cattell uses the term *mental test* in announcing the agenda for his Galtonian test battery.
1901	Clark Wissler discovers that Cattellian "brass instruments" tests have no correlation with college grades.
1905	Binet and Simon invent the first modern intelligence test.
1914	Stern introduces the IQ, or intelligence quotient: the mental age divided by chronological age.
1916	Lewis Terman revises the Binet-Simon scales, publishes the Stanford-Binet. Revisions appear in 1937, 1960, and 1986.
1917	Robert Yerkes spearheads the development of the Army Alpha and Beta examinations used for testing WWI recruits.
1917	Robert Woodworth develops the Personal Data Sheet, the first personality test.
1920	Rorschach Inkblot test published.
1921	Psychological Corporation—the first major test publisher—founded by Cattell, Thorndike, and Woodworth.
1927	First edition of the Strong Vocational Interest Blank published.
1939	Wechsler-Bellevue Intelligence Scale published. Revisions published in 1955, 1981, and 1997.
1942	Minnesota Multiphasic Personality Inventory published.
1949	Wechsler Intelligence Scale for Children published. Revisions published in 1974, 1991.

SUMMARY

1. In 1910, Henry Goddard translated the 1908 Binet-Simon scale. In 1911, he tested more than a thousand schoolchildren with the test, relying upon the original French norms. He was disturbed to find that 3 percent of the sample was "feebleminded" and recommended segregation from society for these children.

2. Nonverbal intelligence tests were invented in the early 1900s to facilitate testing of non-English-speaking immigrants. For example, Knox published a wooden puzzle test in 1914 and also used the now familiar digit-symbol substitution test.

3. In 1916, Lewis Terman released the Stanford-Binet, a revision of the Binet scales. This well-designed and carefully normed test placed intelligence testing on a firm footing once and for all.

4. During WWI, Robert Yerkes headed a team of psychologists who produced the Army Alpha, a verbally loaded group test for average and superior recruits, and the Army Beta, a nonverbal group test for illiterates and non-English-speaking recruits.

5. Early testing pioneers such as C. C. Brigham used results of individual and group intelligence tests to substantiate ethnic differences in intelligence and thereby justify immigration restrictions. Later, some of these testing pioneers disavowed their prior views.

6. Educational testing fell under the purview of the College Entrance Examination Board (CEEB), founded at the turn of the twentieth century. In 1947, the CEEB was replaced by the Educational Testing Service (ETS), which supervised the release of such well-known tests as the Scholastic Aptitude Tests and the Graduate Record Exam.

7. The advent of multiple aptitude test batteries was made possible with the development of factor analysis by L. L. Thurstone and others. Later, the improvement of these test batteries was spurred on by the practical need for selecting WWII recruits for highly specialized positions.

8. Personality testing began with Woodworth's Personal Data Sheet, a simple yes-no checklist of symptoms used to screen WWI recruits for psychoneurosis. Many later inventories, including the popular Minnesota Multiphasic Personality Inventory, borrowed content from the Personal Data Sheet.

9. Projective testing began with the word association technique pioneered by Francis Galton and brought to fruition by C. G. Jung in 1910. Hermann Rorschach published his famous inkblot test in 1921.

10. The Thematic Apperception Test (TAT), a picture storytelling test introduced in 1935 by Morgan and Murray, was based upon the projective hypothesis: When responding to ambiguous or unstructured stimuli, examinees inadvertently disclose their innermost needs, fantasies, and conflicts.

11. The assessment of vocational interest began with Yoakum's Carnegie Interest Inventory developed in 1919–1920. After several revisions and extensions, this instrument emerged as E. K. Strong's Vocational Interest Blank.