A Brief History of Mental Testing.

Bob Lichtenstein, PhD

The field of psychology and the history of mental testing go hand in hand. Psychology had its beginnings in the late 1800s, as scholars turned their attention to measuring human abilities. Wilhelm Wundt in Germany, Francis Galton in England, and James McKeen Cattell and G. Stanley Hall in the United States were among the first to apply scientific methods of psychological measurement. The abilities they chose to study were mostly physiological and sensory. They measured characteristics such as reaction time, visual and auditory discrimination, motor skills, and hand strength under carefully controlled lab conditions.

The modern era of mental testing was set in motion by French psychologist Alfred Binet, who proposed that measurement of human intelligence should focus on more complex and meaningful traits such as language, reasoning, memory, imagination, attention, and comprehension. Binet studied cognitive development much the same way Piaget did 50 years later: he observed what his children were able to do, and not able to do, at various ages. When Binet was selected to serve on a commission for the French government that was charged with improving public education for atypical children, Binet saw this as an opportunity to apply his research to create a test that could identify those children who needed specialized education programs. He and his colleague, Theodore Simon, produced the first version of this test in 1905, and improved upon it with an expanded version in 1908.

The Binet-Simon Scale of Intelligence provided an estimate of a child’s “developmental age,” which could be compared to the child’s chronological age to estimate whether intellectual development was typical, above average, or below average. The IQ, or *intelligence quotient*, originated with a German psychologist who introduced a formula for reducing performance on the test to a single number. This quotient was computed by dividing developmental age by chronological age and multiplying by 100. If the developmental age were equal to the chronological age, the IQ was 100. A ten year old with a developmental age of eight would have an IQ of 80 (8/10 x 100); a ten-year-old with a developmental age of 12 would have an IQ of 120 (12/10 x 100).

Binet was strongly opposed to using his test in this way. He cautioned against perceiving intelligence as a permanent trait that could be represented by a single number and used to rank individuals. He viewed his test simply as a practical means of identifying children who needed special help so they could improve their mental capabilities through education. His concerns went unheeded. Binet was not a highly regarded or influential figure in scholarly circles, and his test was never put to its intended use in France. Instead, the first generation of intelligence tests were used primarily in two ways: to develop theories about the nature of intelligence, and for dubious purposes of social engineering.

The Binet-Simon test received limited attention in Europe, but lit a spark when it crossed the Atlantic Ocean and energized the field of psychology in the U.S. The test was popularized by Henry Goddard, research director of the New Jersey Training School for Feeble-Minded Girls and Boys, who translated it into English and circulated it widely. Goddard proposed a system of labelling individuals based on their IQ tests scores, with the terms *moron, imbecile* and *idiot* referring to successive levels of mental deficiency. Goddard and other American psychologists saw intelligence testing as a way to apply the concept of *eugenics*—improvement of the gene pool—for the ostensible betterment of society. They subscribed to a *hereditarian* theory that regarded intelligence is fixed at birth, as determined by heredity. Individuals with low intelligence were seen, not only as less capable of doing challenging work, but lacking in moral character—lazy at best, criminals at worst, and likely to undermine rather than contribute to the public good.

Louis Terman, a Stanford professor and leading hereditarian, made important refinements to the Binet test. With a first version in 1916 and a significantly expanded version in 1937, his Stanford-Binet Intelligence Scale dominated intelligence testing in the U.S. for half a century. The 1937 version of the Stanford-Binet replaced the quotient method of producing an overall score with a statistical model using the normal (i.e., bell-shaped) curve to statistically compute scores based on one’s performance relative to others in the same age range. The scale was similar, with the population average set at 100; the score continued to be called an IQ, but it was in fact no longer a quotient.

Robert Yerkes, a Harvard professor, saw IQ testing as a way to aid the war effort in World War I. He presided over a mass testing program of American soldiers, with the intention that their scores would determine their military assignments. It soon became clear that the written, group-administered Army Mental Test could not be used with the large number of illiterate soldiers. Yerkes and his collaborators created an alternate version—the Army Beta test, with nonverbal test items such as solving mazes, mentally rotating geometric shapes, and identifying missing the parts of pictured objects. The military made little use of the test results, but they received significant attention after the war. When Yerkes analyzed the massive database from having testing over 1 million servicemen, his research showed that recent immigrants—especially those from countries in Southern and Eastern Europe—obtained lower scores. This was consistent with the findings of Goddard’s earlier study at Ellis Island, from which he concluded that the majority of immigrants from these countries were mentally deficient. Goddard’s study was seriously flawed, since he disregarded the educational, cultural, and language backgrounds of these immigrants. But with the strong support of prominent scholars like Goddard and Terman, the eugenics movement brought about sweeping changes in U.S. policy in the 1920’s that restricted immigration from European countries.

It was Arnold Gesell, a graduate school classmate of Terman’s, who used Binet’s methods for their originally intended purpose. Gesell, a Yale professor and child psychologist who would become a well-known public figure (the Dr. Oz of his day), accepted a special assignment for the Connecticut Department of Education in 1911. His primary activity was essentially what school psychologists do to this day. Gesell used standardized testing and other sources of information to describe the mental functioning and behavior of school-age children and determine their need for special educational services. This work earned him the distinction of being the first school psychologist.

The explosion of mental testing around the beginning of the 20th century provided ample material for Charles Spearman, a prominent British psychologist, to explore the nature of intelligence. The basic question he posed is whether intelligence is a single entity that is expressed in various ways, or a collection of different skills and abilities. Spearman observed that scores on various types of mental tests were positively correlated, meaning, individuals tended to score consistently high, or consistently low, when different mental abilities were measured. He invented factor analysis—a way of analyzing the performance of many individuals on a battery of tests—to see whether intelligence was best represented by a single quantity, which he called “g”, or various specific abilities, or “s” factors. Spearman found clear evidence for general intelligence (“little g”) when he first reported his findings in 1904. His full treatise on the subject in 1927, *The Abilities of Man*, gave credence to the existence of both “g” and “s” factors.

The nature of these various specific abilities—how we divide up intelligence into separate domains, has continued to be a major focus of research by cognitive psychologists to this day. In the 1930’s, Louis Thurstone, an American psychologist, made use of factor analysis to propose that intelligence is made up of seven main types of abilities: numerical, reasoning, spatial, memory, perceptual speed, verbal comprehension, and word fluency. In the 1960s, Raymond Cattell, a British psychologist, added an important dimension by making the distinction between fluid and crystallized intelligence. Cattell saw fluid intelligence as the innate brain power that enabled one to solve new problems and deal with abstract concepts, which he contrasted with crystallized intelligence—the accumulation and application of acquired knowledge that is gained through learning and experience. Cattell’s ideas were further developed by John Horn and John Carroll, and evolved into the current leading theory of intelligence that bears the initials of these three psychologists. *C-H-C theory* breaks down intelligence into a handful of broad factors such as fluid and crystallized intelligence, quantitative knowledge, visual and auditory processing, and short-term memory; and into many narrow factors that contribute to each broad factor.

The Stanford-Binet Intelligence Scale saw its first serious competition with the work of David Wechsler, chief psychologist at New York’s Bellevue Hospital, who in 1939 developed the Wechsler-Bellevue Scale of Intelligence for adults. Wechsler, who was a student of Charles Spearman, regarded intelligence as “the aggregate or global capacity of the individual to act purposefully, to think rationally and to deal effectively with his environment.” Wechsler sampled a range of cognitive functions to capture the richness and complexity of human intelligence, adding nonverbal items like those from the Army Beta test to verbal items like those on the Stanford-Binet. The descendants of the Wechsler Bellevue—the Wechsler Adult Intelligence Scale (WAIS), the Wechsler Intelligence Scale for Children (WISC), and the Wechsler Preschool and Primary Scale of Intelligence (WPPSI)—have become far and away the most widely used tests of mental abilities. The Wechsler tests are ingeniously constructed, enabling straightforward comparisons with national norms and generating a profile of individual strengths and weakness. The tests each consist of about a dozen subtests that yield scores ranging from 1 to 19, with an average of 10 and a standard deviation of 3. Subtests scores are combined to create composite scales with an average of 100 (like the Stanford-Binet IQ) and a standard deviation of 15. Many tests now use this same scoring method, since it is so familiar to test users. The two scales of the original WAIS and WISC—a verbal IQ and a nonverbal IQ—have evolved into a more elaborate structure informed by generations of research. The current version of the Wechsler Intelligence Scale for Children, the WISC-V, is composed of five major scales: Verbal Comprehension, Fluid Reasoning, Visual-Spatial, Working Memory, and Processing Speed.

The psychological testing movement was not limited to tests of mental ability. Projective tests, dating back to the early 1900s, assessed personality traits by having test-takers respond to ambiguous stimuli—describing what Rorschach inkblots look like, or making up stories about the people in the pictures of the Thematic Apperception Test. Projective tests are still widely used, but have been criticized for relying on clinical judgment and for lacking adequate scientific support. Researchers at the University of Minnesota introduced a research-based alternative, an objective personality test called the Minnesota Multiphasic Personality Inventory, or MMPI. The MMPI is a written test consisting of hundreds of true/false items about the test-taker’s thoughts, feelings, and preferences. Many of these items seem odd or irrelevant, but they have shown by actuarial studies to accurately identify psychological traits. Rating scales are now widely used to assess the behavioral and social-emotional characteristics, such as attention, impulsivity, aggression, anxiety, depression, social skills, and daily living skills. Children and adolescents are rated by parents and teachers, or may rate themselves. Rating scales are interpreted objectively, yielding scores that are compared with national norms.

Some approaches to personality testing have fallen by the wayside. The Bender-Gestalt test, a paper-and-pencil test that involves copying a set of designs, was once used as a way to identify emotional problems or brain damage. These applications were shown to be without merit, although the Bender-Gestalt is still used to assess visual-motor skills. In the 1950s and 1960s, when psychoanalysis was in vogue, psychologists developed systems for relating particular types of responses on the Wechsler Adult Intelligence Scale to personality traits, such as paranoia, neuroses, depression, and antisocial tendencies. This approach, too, has been discredited.

Criticisms of intelligence tests peaked in the 1960s and 1970s. Landmark court cases challenged the use of intelligence tests to place minority children and English language learners in separate classrooms, with lower expectations. The Larry P. case levied the charge that intelligence tests were racially biased. The case occupied California courts for a decade, and resulted in a ban on the use of intelligence tests with African-American children that is still in place. Arthur Jensen, a cognitive psychologist at the University of California, fueled the flames with a scholarly paper in 1969 that laid out the scientific case for racial differences in intelligence being due to heredity, rather than environmental differences. Jensen argued that efforts to promote equality through social programs like Head Start were pointless and wasteful. Heated exchanges followed, as did further research and reanalysis of previous studies. The controversy subsided as the scientific community resolved that intelligence is significantly influenced by both heredity and environment.

One consequence of controversies over the use of intelligence tests is that they are now designed with close attention to minimizing bias due to cultural, racial, geographic, and gender differences. Professional organization have developed guidelines to guard against discriminatory use of tests, and professional training programs emphasize understanding of and sensitivity to cultural and language differences. Special education regulations include specific guidelines to ensure that tests are appropriately selected, administered, and interpreted, and do not discriminate on the basis of language differences or sensory impairment.

Nevertheless, language differences remain a particular challenge for test developers. It is virtually impossible to develop norms for English language learners in the U.S., since they have widely ranging language backgrounds—having been exposed to English for varying periods of time and to varying degree. Just translating a test into Spanish is not a simple matter, since there are variations in the Spanish words used in different countries. Nonverbal tests of intelligence have been developed that minimize or even eliminate the use of language. But these tests are limited, because they miss that critically important element—the ability to understand and use language—which has major implications for how children perform in school and how adults perform in many lines of work.

The use of standardized tests for educational purposes has sustained a thriving test publishing industry in the U.S. Group administered tests are used to inform admission decisions to prep school, college, graduate school and professional schools, and to ascertain the competence of teachers, lawyers, physicians, and other professionals. Individually administered tests for children have proliferated since 1975, when Congress passed the special education law now known as IDEA (the Individuals with Disabilities Education Act). Special education laws ensure that children with disabilities receive an individualized education program (or, IEP) based on the results of an educational evaluation. Intelligence tests, achievement tests, and tests of special abilities are the tools commonly used to conduct these evaluations. The need for educational evaluations also created unprecedented demand for special education teachers, school psychologists, school-based speech and language pathologists, and other specialists.

There are dozens of well-developed, nationally normed tests of general intelligence. With each new generation of intelligence tests, test authors place more emphasize on the specific domains they measure, but educators still assign importance to an overall score, or IQ. Some tests of intelligence are designed for specialized purposes, such as screening, or for use with young children, non-English speakers, or individuals with impaired hearing or vision. There are tests that focus on specific aspects of intelligence, such as memory and learning, attention, vocabulary, processing speed, and executive functioning. There are tests for every conceivable purpose: math skills, attention, vocational aptitude, mechanical ability, resilience, social skills, adaptive behavior, and so on. The Mental Measurements Yearbook, the authoritative reference book on published tests, is now in its 19th edition and provides information on over 3000 tests.