THE CAMBRIDGE HANDBOOK OF Intelligence

EDITED BY
Robert J. Sternberg,
Scott Barry Kaufman

CAMBRIDGE
CAMBRIDGE www.cambridge.org/9780521518062

CHAPTER 2

Tests of Intelligence

Susana Urbina

There are many ways of approaching the topic of intelligence tests. This chapter deals with just two of them. One approach centers on what intelligence tests measure and is tied to the issue of defining what intelligence is. The close connection between those two questions can be seen in E. G. Boring's (1923) definition of intelligence as that which intelligence tests measure. Most readers will probably agree that this definition, while easy to remember, is thoroughly unsatisfactory because of its circular nature and limited utility. More substantial and satisfying definitions can be found later in this chapter and in many other sources (e.g., Sternberg & Detterman, 1986; Urbina, 1993). Boring's definition, such as it is, does provide us with a reason to examine what the multiplicity of intelligence tests do measure and thus understand what some of the basic aspects of the construct of intelligence are, at least in the cultures that gave rise to those tests.

The second way to approach the topic of intelligence tests is far more pragmatic. It concerns the issue of why these tests exist or the purposes for which they are employed. In an interesting but not altogether surprising coincidence, both ways of approaching intelligence tests – clarifying what they measure and what kinds of practical purposes they can serve – date back to the beginning of the 20th century.

This chapter reviews the basic elements of both approaches by examining intelligence tests in some detail. In particular, it poses and attempts to answer the following questions:

- What are intelligence tests?
- When and how did intelligence tests come to be?
- Do intelligence tests really measure intelligence?
- What do intelligence tests actually do?
- What functions or purposes do intelligence tests serve?
- Do intelligence tests have a future?

What Are Intelligence Tests?

The latest edition of the *Tests in Print* (TIP) series (Murphy, Spies, & Plake, 2006) lists

202 tests in the "Intelligence and General Aptitude" category. Of these, only 27 tests use the term *intelligence* in their titles. This number has not changed since the previous edition of TIP. By and large, the tests published in the past few decades avoid using intelligence in their titles, whereas the older tests continue to do so, even in their new editions, in order to provide continuity and because their names are well established.[1] In addition, the traditional intelligence tests – especially the Wechsler scales and the Stanford-Binet–also have been the most widely used and studied (Camara, Nathan, & Puente, 2000). If one examines the items and manuals of the tests within the TIP category of "Intelligence and General Aptitude," one finds striking similarities of both form and purpose among them, whether or not they have the word intelligence in their titles.

The truth about IQ tests. Although the phrase "IQ test" is frequently used to refer to intelligence tests, the two terms are not at all equivalent. The confusion between them stems from the fact that the earliest intelligence tests, such as the Stanford-Binet, used a score called the *intelligence quotient* or IQ for short. Originally, the IQ was an actual quotient obtained by dividing a number labeled Mental Age (MA) – which reflected a person's performance on the test and was expressed in years and months – by the person's Chronological Age (CA) and multiplying the result by 100 to eliminate the decimals. If performance on the test or MA matched the person's CA exactly, the IQ would be 100. Hence that number became known as the "normal" or average intelligence level. Numbers above and below 100 indicated that performance on the test had exceeded or fallen short of the levels expected at a given CA and became associated with above and below average intelligence, respectively. Eventually it became clear that, for a variety of reasons, this way of obtaining intelligence test scores did not work well – especially in adulthood when mental development levels off so that increases in CA cannot be matched by corresponding increases in MA. Thus, a new way of arriving at IQ scores was devised.[2]

The newer measure, known as the *deviation IQ*, is the type of score currently in use by the major tests that still use the IQ. In spite of the label, the deviation IQ is no longer a quotient. Instead, IQs are now derived by comparing a person's performance or raw score on a test of intellectual abilities to norms established by the performance of a representative group – known as a *normative* or *standardization sample* – of people in the person's age range. Raw scores for each normative age group are converted into standard scores with a mean of 100 and a standard deviation (SD) typically set at 15. The difference between a person's score and the average score of her or his age group – in SD units – determines the person's IQ. Thus, deviation IQ scores of 85 and 115 are 1 SD unit away from the mean and both reflect performance that deviates equally from the average performance of a comparable age group sample, but in opposite directions. Since test scores obtained from representative samples produce distributions resembling the normal curve model, they can be made to fit into the normal curve parameters so that *approximately* 68% of the scores are within ±1 SD from the average, 95% are within ±2 SD, and 99% are within ±3 SD. This is just one of the reasons to be suspicious of reported IQ scores much higher than 160, which – if the SD is set at 15 – is a number that would represent performance at 4 SDs above the average and thus in the top one-tenth of 1% of the age group norm. IQ scores much higher than 160 cannot be obtained in most of the current tests of this type.

As of now, the TIP lists barely more than a dozen tests that produce IQ scores. These include the current versions of the oldest traditional intelligence test batteries,

1 Tests within the cited TIP category that were published since the 1970s or 1980s tend to use terms such as *cognitive abilities*, *general ability*, or simply *aptitude* in their titles.

2 For a more complete history of the IQ score, see Murdoch (2007).

such as the Stanford-Binet Intelligence Scale (SB), the Slosson Full-Range Intelligence Test (S-FRIT), the Wechsler Adult Intelligence Scale (WAIS), the Wechsler Intelligence Scale for Children (WISC), and the Wechsler Preschool and Primary Scale of Intelligence (WPPSI). Some test batteries of more recent vintage also yield IQ scores, notably the Kaufman Adolescent and Adult Intelligence Test (KAIT), but most of the newly developed tests that yield IQ scores are either abbreviated versions of other tests, such as the Wechsler Abbreviated Scale of Intelligence (WASI) and the Kaufman Brief Intelligence Test (K-BIT), or tests limited to nonverbal content, such as the Universal Nonverbal Intelligence Test (UNIT), the Leiter International Performance Scale-Revised (Leiter-R), or the General Ability Measure of Adults (GAMA). Due to the controversies surrounding IQ scores and to the excessive and unjustified meanings that the IQ label has acquired, the use of IQs in scoring intelligence or general aptitude tests is rapidly being abandoned, replaced by terms such as *General Ability Score* or *Standard Age Score*. In keeping with tradition, however, most of these scores are derived in the same way as deviation IQs and have a mean set at 100 and SDs of 15 or 16.

When and How Did Intelligence Tests Come to Be?

The origins of intelligence testing are inextricably linked to Francis Galton and Alfred Binet. Of course there were others – both before and after them – who contributed to the development of intelligence tests in significant ways, but these two men, who had very different goals, set the stage for most of the positive and negative consequences that would follow. Accounts of the history of intelligence testing and of the leading figures in that history, as well as of the controversies they generated, can be found in many sources. Among the most interesting and readable ones are those provided by Fancher (1985), Sokal (1987), and Zenderland (1998).

Among psychologists, Francis Galton is most often remembered as the originator of the so-called "nature-nurture" controversy that has been such a crucial point of debate in the social sciences. Galton's desire to devise a way to measure intelligence stemmed from his interest in giftedness and genius and his eugenicist notion that the intellectual caliber of society would be improved by identifying highly intelligent young men and women and encouraging them to procreate early and profusely. This idea, in turn, arose from his conviction that intelligence is an inherited and unitary trait rooted in physiology. Using the theory of evolution developed by his cousin Charles Darwin as a source of inspiration, Galton investigated the extent of resemblance in terms of intellectual achievement among people with different degrees of familial ties. Even though his findings were insufficient to prove his argument conclusively, Galton nevertheless proceeded to develop a series of measures of reaction time, sensory acuity, and such, which he believed were indices of one's natural inherited ability associated with functions of the central nervous system. Although Galton collected such data on thousands of individuals at his Anthropometric Laboratory in England, it was left to an American psychologist named James McKeen Cattell – who was influenced by Galton – to continue this line of work in the United States and to see the premises on which it was based discredited. Cattell coined the term *mental tests* to refer to a series of tasks involving primarily psychomotor and sensory measures along the lines of those suggested by Galton's theory and he proceeded to collect data using these measures at Columbia University. Unfortunately for the theory, a study by one of Cattell's own students (Wissler, 1901) indicated that there was practically no relationship among the mental tests or between them and the indices of academic achievement used as a criterion of mental ability.

Whereas Galton, as well as Cattell, failed in his endeavor to create a device for assessing intellectual abilities, their French

contemporary Alfred Binet succeeded admirably. Unlike Galton, Binet worked with children and was interested in identifying intellectual retardation rather than giftedness. He got involved in this effort in 1904 when he was appointed by the French government to a commission whose task was to implement the new law requiring public education for all children. Identifying individuals who, due to mental retardation, would be unable to attend ordinary schools and would require special education was an essential aspect of this mandate. Due to a variety of circumstances in his personal and professional life, Binet was at that point particularly well prepared for the job he undertook (Wolf, 1973). He and his collaborator Theodore Simon were able, by 1905, to develop and publish a scale consisting of 30 simple tasks of increasing difficulty that could distinguish among children with different levels of intellectual capacity. Binet and Simon used their experiences with this first scale to extend and refine it, concentrating on those items that had proved most useful in discriminating among children of different ages and mental capacity levels. They realized that by tapping a variety of cognitive tasks – such as memory, attention, verbal comprehension, and reasoning – at different levels of difficulty and organizing the items according to the age levels at which children of normal intellectual functioning were likely to succeed, they could produce a scale that would classify children's levels of mental functioning based on the number of items they passed at the various levels. In 1908 and 1911 Binet and Simon published considerably improved revisions of their scale, which quickly gained in popularity, especially in the United States where the scales were almost immediately translated, used, and distributed at the Training School for the Feebleminded in Vineland, New Jersey, by its director of research, Henry H. Goddard.

In fact, after Binet's death in 1911, the main center of research and test development on intelligence shifted from Europe to the United States where several other adaptations of the Binet-Simon scale were being tried out, culminating with the publication, in 1916, of the Stanford Revision of the Binet-Simon Intelligence Scale developed by Lewis Terman and his graduate students at Stanford University. This scale, which became known as the Stanford-Binet (SB), was considerably expanded and was adapted for and standardized on children from the United States. In addition, Terman decided to use the IQ formula – MA/CA times 100 – to express scores on the SB scale. In spite of the fact that the SB was primarily suitable for children, this scale dominated the field of individual intelligence testing for the next few decades. The SB was singularly responsible for popularizing the IQ score, which became synonymous with intelligence and was adopted by several other tests of abilities, some of which are still in use today. In fact, when David Wechsler published each of his series of enormously successful intelligence tests, starting in 1939 with the Wechsler-Bellevue Intelligence Scale, he chose to keep the term *IQ* to designate the scores on those scales. As mentioned earlier, Wechsler's deviation IQs, were very different from the SB IQs in that they were no longer quotients and could be meaningfully applied to people of all age groups.

Group intelligence tests. Whereas Binet and Wechsler are famous for their overwhelming impact on the field of individual intelligence tests, the person most responsible for the development of group tests, Arthur S. Otis, is not as well known. Otis studied with Lewis Terman at Stanford University in the years prior to World War I and became intrigued by the possibility of adapting some of the tasks of the Binet scale for use with groups in a paper-and-pencil test format. One of the most significant innovations that Otis devised was the multiple-choice type of item format. This innovation, in turn, was instrumental in the development of the first group test of mental ability, namely, the Army's Group Examination Alpha also known as the Army Alpha, which was used in the selection and classification of Army personnel during the First World War.

The success of the Army Alpha spawned the rapid development of many other paper-and-pencil tests of cognitive abilities. Otis himself developed the Otis Group Intelligence Scale, published in 1918, which was the first American group test of mental ability specifically designed for use in educational institutions. Otis developed other tests of mental ability and contributed several innovations and refinements that made the scoring and administration of group tests more practical and efficient (Robertson, 1972). The Otis-Lennon School Ability Test, Eighth Edition (OLSAT8), which is the current version of the Group Intelligence Scale, is still widely used to evaluate cognitive abilities related to success in school from kindergarten to 12th grade. Another contemporary group test designed for the same purpose and population is the Cognitive Abilities Test, Form 6 (CogAT-6). At the higher education level, the College Board's SAT Reasoning Test and the Graduate Record Examination General Test are the prime examples of group tests used to screen applicants in terms of their level of cognitive abilities.

In addition to the Army Alpha, which no longer is used, a variety of other group tests have been developed and used – though not always wisely or effectively – by military and civilian organizations to select and classify personnel. Some of these tests, such as the Wonderlic Personnel Test (WPT) – originally adapted from the Otis Self-Administering Tests of Mental Ability – attempt to get a general estimate of cognitive ability, whereas others are aimed at evaluating specific skills required for performance in a given occupation, such as clerical or mechanical abilities.

Do Intelligence Tests Really Measure Intelligence?

The short and simple answer to this question is no. Given that semantics play a large part in this answer, a review of the meaning of the terms in the question may clarify the answer. The meaning of *measure* is clear: to measure something is to assign numbers or labels to objects, events, or people according to some established method or rules (see Kirk, 1999, e.g.). Based on this definition, we can establish that intelligence tests do measure something. After all, they produce numbers that are assigned to the responses of test takers on the behavior samples that make up each test, and those numbers are assigned according to designated standards or rules.

Whether what intelligence tests measure is intelligence, on the other hand, is far more complicated as even a casual perusal of the field should reveal. Although many people assume that since intelligence tests exist, it must be possible for intelligence to be measured, the fact is that intelligence is an abstraction, a construct we infer based on the data at our disposal and our own criteria. As such, it is not something everyone can agree on or quantify objectively.[3] Thus, even among psychologists there is a wide variety of opinion about the meaning of intelligence, depending on the perspective from which they approach the topic.

Neither Galton nor Binet ever really defined intelligence. In fact, Galton seldom even used the term. Nevertheless, Galton's observations led him to believe that intelligence or general mental ability is a single hereditary, biological trait that is largely responsible for outstanding achievements in any field of endeavor. Although he recognized the existence of additional special aptitudes for certain fields, such as music and art, Galton believed that in order for these abilities to reach expression in extraordinary accomplishments, they had to be paired with an innate and superior level of general ability (Jensen, 1998).

The closest Binet came to defining intelligence was in an article he co-authored with

3 One of the many reasons the question of which of the two sexes is more intelligent cannot be answered is that most intelligence tests are deliberately constructed in a way that will result in no overall sex difference by balancing tasks that favor females and those that favor males.

Simon (1904) in which they equate intelligence with judgment or common sense, adding that "to judge well, to comprehend well, to reason well" (p. 197) are the essential *activities* of intelligence. Unlike Galton, Binet believed that intelligence consists of a complex set of abilities – such as attention, memory, and reasoning – that are fluid and shaped by environmental and cultural influences. Binet was also far less inclined than Galton to believe that intelligence could be reliably or precisely measured. He thought that to the extent that his scale captured some of the essential aspects of intellectual functioning, it would prove more serviceable in evaluating those at the subnormal range rather than at the superior levels of intellectual functioning that were Galton's primary concern.

Although it was Binet who succeeded in producing a practical method for estimating mental ability and in providing a useful solution to the problem of identifying children at the lower end of the ability spectrum, his notions about the nature of what his method was actually tapping were not, by any means, universally adopted. On the contrary, Binet's successful technique and the great variety of tests that proliferated following his lead provided additional means for other investigators to carry on research programs influenced by Galton's ideas. In particular, Charles Spearman's application of factor analysis to data derived from mental tests led him to believe that though numerous *specific* (*s*) factors are involved in the performance of tasks requiring specialized abilities, there is an overarching *general* (*g*) factor that is implicated to a greater or lesser extent in *all* intellectual activities (Spearman, 1927). Although Spearman himself thought of the *g* factor as a mathematical abstraction and did not equate it with intelligence, many others did and continue to do so (see, e.g., Gottfredson, 2009). In opposition to this, other theorists propagated views that were more in line with Binet's. L. L. Thurstone, for example, also applied factor analytic techniques to mental test data but, unlike Spearman, he argued that there are several distinct and *independent* group factors, such as verbal comprehension, numerical reasoning, memory, and such involved in intellectual activities (Thurstone, 1934). Much of the disagreement between those who supported Spearman's emphasis on the singular role of the *g* factor and those who favored multiple factors was based on different ways of conducting factor analyses on ability test data, as well as on the number and types of tests included in the analyses.

Aside from Binet, the other towering figure in the history of intelligence testing is David Wechsler. The test series that Wechsler developed starting in the 1930s, much like the scales originated by Binet in an earlier time, became the most widely used instruments for the individual assessment of intelligence and have been, for several decades, the standard against which other such tests are compared. Unlike Binet, however, Wechsler did provide a carefully crafted definition of intelligence which he modified somewhat over time. In the final version of that definition, Wechsler stated that intelligence is "the aggregate or global capacity of the individual to act purposefully, to think rationally and to deal effectively with his [*sic*] environment" (1958, p. 7).

Wechsler studied with Cattell and Spearman as well as with E. L. Thorndike, a psychologist whose views of intelligence differed considerably from Spearman's. Based on this training, he developed a position on intelligence that encompassed aspects of each of their viewpoints. In addition, Wechsler had been directly involved in administering and helping to develop intelligence tests since the time of World War I. As a result, when he started his own work on test development, Wechsler was uniquely qualified to address the topic of intelligence and its measurement. Near the end of his life, hoping to facilitate consensus about how to assess intelligence, Wechsler (1975) wrote an article in which he clearly aimed to debunk some of the common assumptions about the nature and meaning of intelligence that had led to the many conflicting views

of it. Among the more interesting points Wechsler made in this article, were the following:

- intelligence is not a quality of mind, but an aspect of behavior;
- intelligence can neither be defined in absolute terms nor equated with cognitive ability;
- intelligent behavior requires nonintellectual capabilities, such as drive and persistence, as well as the ability to perceive and respond to social and aesthetic values; and
- intelligent behavior must not only be rational and purposeful; it must also be esteemed.

In this article, Wechsler quite sensibly admitted that intelligence is a relative concept. When it comes to intelligence tests, Wechsler stated his belief that they are valid and useful and that a competent examiner can do much better at evaluating intelligence with them than without them. Considering that he was keenly aware that his reputation would rest on the intelligence scales bearing his name, this is not surprising. In the final paragraph of the article, however, Wechsler came up with this puzzling conclusion:

> *What we measure with tests is not what tests measure – not information, not spatial perception, not reasoning ability. These are only means to an end. What intelligence tests measure, what we hope they measure, is something much more important: the capacity of an individual to understand the world about him and his resourcefulness to cope with its challenges. (Wechsler, 1975, p. 139)*

Such a conclusion might be tenable if Wechsler had said that intelligence tests allow us to *infer* an individual's capacity to understand the world and to cope with its challenges. However, as stated, his conclusion is puzzling in that it negates the possibility that tests measure some fairly well-defined and clear-cut constructs while suggesting that they can measure an infinitely more complex one. For who can doubt that what Wechsler meant by "the capacity . . . to understand the world" and the "resourcefulness to cope with its challenges" was anything other than intelligence itself?

What Do Intelligence Tests Actually Do?

Notwithstanding Wechsler, all intelligence tests – indeed all psychological tests of any kind – measure nothing more or less than samples of behavior. In the case of intelligence tests, the behavior samples are relevant to cognitive abilities of one sort or another and these abilities, in turn, have a very significant impact in various life outcomes, such as educational and occupational success. For example, many intelligence tests sample test takers' knowledge of vocabulary by asking them to define words at various difficulty levels, ranging from simple words used in everyday speech to more difficult and obscure ones. Test takers' scores depend on the number and difficulty of the words they are able to define and on how well that compares to what others in their age group can do. To a large extent, performance on vocabulary tests depends on the amount of reading people do and – all other things being equal – people who read more tend to acquire a larger fund of knowledge, understand verbal communications better, and do better in academic work than people who read less. Thus, while all that is measured by a vocabulary test – provided the words have been correctly scaled in terms of difficulty and provided the age group used for comparison is appropriate – is the level of a test taker's vocabulary compared to her or his age peers, what we can *infer* based on that measure is much more than that. Intelligence tests rely for their validity on the demonstrable relationships between the samples of behavior they tap and what can be justifiably inferred from those samples in terms of general ability. In addition to vocabulary, which is typically a reliable indicator of a person's general intellectual ability, intelligence tests include behavior samples

that require quantitative, verbal, and visual-spatial reasoning skills as well as processing speed and various kinds of memory.

The question of validity. If we agree with Wechsler's argument, reiterated by Anne Anastasi years later, that "intelligence is . . . a quality of behavior" and that intelligent behavior is displayed in "effective ways of coping with the demands of a changing environment" (Anastasi, 1986, pp. 19–20), it follows that intelligence cannot be measured or encompassed by a single number. Nevertheless, for approximately the first half of the 20th century, from the time of the original Binet-Simon scales until the Wechsler scales for adults and children took over the preeminent role in intelligence testing, many – if not most – psychologists and educators as well as the general public assumed that the IQ was just such a number. This erroneous assumption was due in part to the enormous influence of the Stanford-Binet, which for much of its history yielded a single global IQ score that generally seemed to correctly classify people at the extreme levels of intellectual functioning. Unfortunately, however, this led to a proliferation of so-called "IQ tests" and to some egregious misuses which have been pointed out by critics from several perspectives throughout the history of these instruments (see, e.g., Gould, 1996; Stanovich, 2009).

In spite of the oftentimes virulent critiques to which intelligence tests have been subjected as a result of their misapplications, several of the traditional ones, such as the Stanford-Binet and Wechsler scales, continue to be used and new ones continue to arise. Furthermore, as discussed in a later section, the older scales have been repeatedly revised – and improved – as they have confronted new generations of instruments that apply advances from cognitive and psychometric theory in their development. A good part of the continued popularity of intelligence tests is due to the renewed ascendance of Spearman's notion of *g*. This, in turn, results from the accumulation of decades of factor analytic research confirming the existence of a theoretical construct that accounts for a large portion of the variance in the performance of intellectual tasks, namely, the *g* factor (Carroll, 1993; Jensen, 1998). Although it must not be assumed that the *g* factor and intelligence are the same, or that an IQ score is a direct measure of *g*, the major comprehensive intelligence test batteries are made up of subtests which, for the most part, have high loadings on *g*, as shown by factor analyses of their intercorrelations. In addition to the findings of numerous factor analytic studies, the major arguments for the validity of intelligence tests are based on (a) their high levels of reliability, as demonstrated by internal consistency and temporal stability coefficients that are typically in the .90s range for the total scores and global indices; (b) the extremely high correlations – in the .80s and .90s range – between the global scores produced by most of the major intelligence tests; and (c) the marked differences in the scores that various special populations, such as individuals with different levels of mental retardation or various learning disabilities, obtain (see, e.g., Flanagan & Harrison, 2005; Kaufman & Lichtenberger, 2006).

The latest version of the *Testing Standards* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) defines validity as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). With this definition, the burden of determining whether a particular application of intelligence test scores is valid is placed entirely on the person or institution responsible for the selection and administration of the test, for the interpretation of the scores, and for any decisions or actions taken on the basis of those scores.

Varieties of intelligence tests. There are, at least, four basic ways in which intelligence tests may be classified: (a) by administration mode, that is, individual versus group tests; (b) by the population for which they are intended, such as tests aimed at children or adults, or at other specific groups; (c) by type of content, such as verbal and nonverbal tests; and (d) by whether they are

full-length batteries or abbreviated versions. Although this classification of tests is based on those that carry the term intelligence in their title, it could just as well apply to those that use different labels, such as general or cognitive ability tests.

A thorough discussion of all the varieties of intelligence tests is beyond the scope of this chapter. Nevertheless, a few critical points about these distinctions are necessary in order to understand the field even in the most general terms.

Mode of administration. Individual tests are those administered one-on-one, by a highly trained examiner to a single examinee. The need for thorough training of examiners is critical in this type of test administration because the procedures for presenting items, scoring responses, and handling the test stimulus materials and timing the tasks need to be strictly followed to comply with standardization requirements. When tests of this type are properly used, they provide the examiner with the opportunity to observe the examinee in the process of responding to challenging tasks presented in a highly structured format that is uniform for all examinees. Thus, in addition to scores, these tests yield a wealth of information that can prove extremely useful in clinical assessment. By the same token, it follows that when individual tests are not administered or scored according to standardized procedures, the reliability of results obtained comes into question. *Group tests*, on the other hand, can be administered safely to large numbers of people by almost anyone familiar with some very simple procedures and can be scored objectively. Thus, what is lost in terms of the type of information that can be gathered about the test taker with individual tests is made up in terms of efficiency and economy by group tests. Which type of test should be used depends on the purpose of the assessment and the available resources with which to do it.

Target population. The population for whom tests are intended is critical in at least two ways. It is crucial to remember that all normative scores, such as deviation IQs, indicate only the position or rank of a person's performance when compared to the specific group of individuals who comprise the norms for the test, not how intelligent a person is in any more basic sense. For example, if a test is to be used with adults over the age of 70, it is important to know if normative data were gathered from individuals who represent that population adequately, not only in terms of age and demographic characteristics but also with regard to variables such as living arrangements and health status. Average performance gauged in comparison to institutionalized older adults in nursing homes would be very different from average performance compared to people of the same age living independently.

The Flynn effect. The relative nature of the normative scores employed by intelligence tests is pointedly exemplified by the so-called Flynn effect. Starting in the 1980s, Flynn (1984, 1987) documented a trend that was interpreted as a general rise in the IQ of populations based on the observation that when tests like the Wechsler scales and the Raven's Progressive Matrices Test were revised and updated, successive normative samples set higher standards of performance than the groups employed in earlier versions. Naturally, this finding gave rise to questions regarding the possible reasons for this phenomenon as well as questions about why intelligence test performance would be rising while scores on tests such as the SAT, as well as other indices of academic achievement were not (Neisser, 1998). The changes that Flynn noted have been attributed to a variety of biological and environmental causes – such as better nutrition, medical advances, technological developments, and familiarity with the types of items of intelligence tests – but have never been satisfactorily explained. In fact, some studies have pointed out that the trend for ever-increasing standards in intelligence test performance is slowing or even reversing, at least in developed countries (Sundet, Barlaug, & Torjussen, 2004; Teasdale & Owen, 2005). Regardless of what cause(s) may be

responsible for the fluctuations in intelligence test scores known as the Flynn effect, it is clear that they reflect *relative* changes in the performance of people from different generations on some of the cognitive abilities that the intelligence tests assess rather than in the more comprehensive view of intelligence as a quality of behavior that allows individuals to cope effectively with their environment. In particular, the rise in intelligence test performance standards is more pronounced in tasks that demand *fluid intelligence*, which involves the processing of new information and the solution of novel types of problems, as opposed to those that require *crystallized intelligence*, which entails the application of consolidated knowledge typically acquired in academic settings (Horn & Cattell, 1966).

Test content. The Flynn effect highlights another aspect of intelligence tests that has important consequences for their results, namely, the content of the tests. The most obvious distinction in this regard is between verbal and nonverbal test content, that is, between tests that require the use of receptive and expressive language and those that do not. In general, nonverbal tests of abilities, such as the Raven's Progressive Matrices and the Performance subtests of the Wechsler scales, rely on figural stimuli and visual-spatial reasoning tasks and tend to show larger gains in performance across successive generations than tests that rely on language (Flynn, 1987). Nonverbal tests also are generally considered to be less susceptible to the influence of culture. The verbal-nonverbal test content distinction has an impact both in deciding which type of test is appropriate for a given population and in determining the meaning and significance of test results. Nonverbal tests have been used with ethnically, linguistically, or otherwise culturally diverse populations based on the premise that by removing the influence of language such tests are less culture-laden and thus fairer. By instituting this limitation in content, however, the nature of the construct that is assessed may also be limited and the capacity of intelligence test scores to predict future performance in many academic or occupational endeavors that require verbal abilities may consequently be reduced.

Test length. A similar caveat, in terms of interpretability, applies to intelligence tests that differ in length from their original prototypes, such as the WASI or the K-BIT, which are short tests from the Wechsler and Kaufman series, respectively. When validity information for such brief tests is presented in the form of very high and positive correlations with longer versions or with each other, it simply means that the rank order positions of test takers' scores on both tests is substantially the same. High as those validity coefficients may be, however, they clearly do not mean that the results of the shorter tests are comparable to those of the full batteries either in terms of the range of abilities they tap or in the amount of information about a person's cognitive functioning they provide. See Homack and Reynolds's (2007) *Essentials of Assessment with Brief Intelligence Tests* for a useful and compact introduction to the subject featuring four of the most prominent examples of this type of instrument.

What Functions or Purpos Do Intelligence Tests Serve?

For the purpose of the discussion that follows, the term *intelligence tests* refers only to the full-length comprehensive batteries – based on large and representative samples of children or adults in the United States population – that are individually administered, regardless of whether their titles include the word intelligence. The major current examples of this type of test batteries – besides the Stanford-Binet, Fifth Edition (SB5; Roid, 2003) and the Wechsler scales (WAIS-IV, WISC-IV, & WPPSI-III; Wechsler, 2008, 2003, 2002) – are the Cognitive Assessment System (CAS; Naglieri & Das, 1997), the Differential Ability Scales (DAS-II: Elliott, 2007), the Kaufman Adolescent and Adult Intelligence Scale

(KAIT; Kaufman & Kaufman, 1993), the Kaufman Assessment Battery for Children, Second Edition (KABC-II; Kaufman & Kaufman, 2004), the Reynolds Intellectual Assessment Scales (RIAS; Reynolds & Kamphaus, 2003), and the Woodcock-Johnson III Test of Cognitive Abilities (WJ III; Woodcock, McGrew, & Mather, 2001). Although some group tests, brief tests, or tests that sample only nonverbal content are often used for the same purposes as the comprehensive intelligence tests, their limitations in length, content, or mode of administration are such that they cannot provide the same wealth of information that intelligence test batteries do.

The impact that intelligence tests have had on both the professional and lay notions of what intelligence is, and on the almost complete identification of intelligence with the IQ score, cannot be overestimated. In order to understand this, it helps to review the makeup of those tests, starting with the Stanford-Binet. From the beginning, the Binet scales were age-based in their organization and in the way their results were interpreted. As Binet figured out, by including items in his scale that tapped a variety of cognitive functions – such as verbal comprehension, logical reasoning, and memory – at different levels of difficulty, he could assess children's levels of mental development. So for the better part of its history, until the Stanford-Binet, Fourth Edition, was published (Thorndike, Hagen, & Sattler, 1986), the Binet scales were organized according to age levels, with a heterogeneous mixture of item types for each chronological age level covered by the scales. Thus, the examiner first had to establish a basal age; this was the age level at which all items were passed and before the level at which the first failure occurred. To begin testing, the examiner estimated the age level at which the examinee was likely to succeed with some effort, based on the examinee's chronological age and background. The examiner would then proceed by administering all of the various types of items designated for that age level. At the younger age levels, appropriate for preschool children, items would include simple performance tasks, such as stringing beads, sorting buttons, or tying knots as well as some verbal tasks such as naming objects or repeating series of two or three digits. As the age levels progressed, items would naturally be more difficult and would rely heavily on verbal comprehension and reasoning tasks, such as word definitions and explaining the meaning of proverbs. Depending on how many items were passed at levels subsequent to the basal age, testing would continue until a ceiling age was reached. The procedures for establishing a basal and a ceiling age were quite important as it was critical to determine reliably the age level below which it could be safely assumed that all items would be passed (basal age) or above which all further items would be failed (ceiling age). The mental age (MA) score on the SB was obtained by adding to the basal age credit in years and months for the items the examinee had passed above her or his basal age. Although the specific bases for determining the SB IQ varied somewhat over time, until the fourth edition, the IQ score hinged on the relationship between the MA and the CA of the examinee.

The advent of the Wechsler scales brought many changes that would have significant consequences for the way in which intelligence is assessed. Most of these changes stemmed from the fact that Wechsler intended to develop an instrument suitable for adults. As a result, Wechsler adopted the use of a point scale, rather than an age scale like the one employed by the SB. Thus, in all of the Wechsler intelligence scales, starting with the original Wechsler-Bellevue, items of the same type are arranged in order of difficulty and organized into 10 or more subtests of homogeneous content. Examinees are presented with one subtest at a time and earn points based on how many items they pass on each subtest. In addition, subtest scores can be grouped in a variety of ways. The traditional Verbal and Performance subscale categories, for example, grouped subtests based on whether their content was primarily verbal or not. Subtests such as Information, Vocabulary, Comprehension, and

Similarities made up the Verbal subscale whereas Block Design, Picture Completion, Picture Arrangement, and Object Assembly were among the subtests making up the Performance subscale. The Wechsler scales originally yielded Verbal and Performance IQs (VIQs and PIQs), based on the respective subscales, as well as a Full Scale IQ (FSIQ) based on a combination of the full range of subtest scores.[4] More recently, subtests have been grouped into index scores – namely, Verbal Comprehension, Perceptual Reasoning, Working Memory, and Processing Speed – that are empirically derived on the basis of factor analyses of subtest data. As mentioned earlier, Wechsler also adopted and popularized the use of deviation IQs based on the extent to which examinees' raw scores differ from the mean of their corresponding age group in the standardization sample. Because one's performance is compared to that of the most closely similar age group, IQs obtained in this fashion make sense in that they indicate whether that performance is at, above, or below average – regardless of the age of the examinee.

Even though, from the beginning, the Wechsler scales produced scores on a variety of subtests besides the IQs, for most practical purposes their interpretation was limited to classifying test takers in terms of their general level of intellectual functioning, based on the FSIQ. As time went by, however, the Wechsler scales acquired an overwhelming popularity compared to the SB, especially among clinical psychologists who realized that the variety of scores the Wechsler scales yielded afforded the opportunity to develop diagnostically significant interpretive hypothesis based on particular aspects of an examinee's performance. For example, according to traditional theories of brain organization – which aligned the left hemisphere with language functions and the right hemisphere with spatial skills – differences in the Wechsler Verbal IQ (VIQ) and Performance IQ (PIQ), if present and sufficiently large, were interpreted as indications of dysfunction in either the left or right cerebral hemispheres, depending on whether the PIQ was larger than the VIQ or vice versa. An excellent summary of the research on neuropsychological correlates of VIQ-PIQ discrepancies provided by Kaufman and Lichtenberger (2006), however, leads to the conclusion that whereas right hemisphere and bilateral brain damage often is reflected in a VIQ>PIQ pattern, left hemisphere damage does not show a PIQ>VIQ discrepancy consistently enough to be of diagnostic benefit.

The practice of analyzing the pattern of responses to items and subtests of the Wechsler scales to extract information about test takers' cognitive abilities and psychological functioning beyond that provided by a single summary score was given impetus by Rapaport, Gill, and Schafer (1945, 1946) who proposed a system that was adopted by many psychologists and was augmented over the next few decades. This practice, which became known as profile analysis, was largely based on the observations of clinicians and their experiences with various types of patients. By the 1990s, profile analysis of Wechsler subtest data came under serious criticism, notably by McDermott, Fantuzzo, and Glutting (1990) who pointed out that such analyses as commonly applied for diagnostic purposes suffered from inadequate reliability and validity data and could thus lead to too many incorrect inferences.

Even before disagreement with the traditional ways of analyzing and interpreting intelligence test score profiles was voiced, there were indications of dissatisfaction with the Stanford-Binet and Wechsler scales. This dissatisfaction stemmed from two sources. One was the increasing emphasis the testing professions started to place on the need for multiple sources of validity evidence (see, e.g., American Psychological Association, 1974; American Educational Research Association, American Psychological Association, & National Council on Measurement

4 Verbal and Performance IQs have been abandoned in favor of index scores in all the current versions of the Wechsler intelligence scales except for the WPPSI-III.

in Education, 1985). In this regard, for example, it now seems remarkable that the manual for the WISC, published in 1949, did not mention validity at all and even the WAIS-R, published in 1981, dealt with the topic in three short paragraphs, basically asserting that the validity of the WAIS-R stemmed from its close connection with the Wechsler-Bellevue, which in turn was correlated with other intelligence tests of that time. Thus, over time, simply demonstrating that the scores on intelligence tests were highly correlated with each other came to be perceived as a clearly insufficient basis for establishing their validity for diagnostic purposes.

Another significant source of discontent with the Binet and Wechsler scales stemmed from the fact that theories of intelligence had continued to evolve in the decades following the creation of those tests. One of the main driving forces in the theorizing about intelligence was the continuous and voluminous accumulation of factor analytic research on human cognitive abilities, best summarized by Carroll's (1993) encyclopedic survey of studies on that topic. This research, in turn, led to a useful model of cognitive trait organization.

As a consequence of the changes just described, simple global estimates of general ability or *g*, while useful in projecting the likelihood of success in academic and job settings (see, e.g., Neisser et al., 1996), were increasingly seen as not providing enough clinically useful information about a person's cognitive functioning to justify the cost and time involved in the administration, scoring, and interpretation of a full-length comprehensive individual intelligence test. Furthermore, as theoretical views of intelligence evolved, and advances in neuroscience provided new information about the role of the brain in cognition, it became clear that the comprehensive instruments for the assessment of cognitive abilities could and should be grounded on these more firm theoretical and empirical bases.

One of the first significant steps in the development of a new generation of intelligence tests was the publication of the Kaufman Assessment Battery for Children (K-ABC; Kaufman & Kaufman, 1983). In developing this instrument, Alan and Nadine Kaufman used the differentiation between sequential and simultaneous types of cognitive processing, based on the theories of the Russian neuropsychologist A. R. Luria, as one of the organizing principles in their battery. Prior to developing the K-ABC, Alan Kaufman – who had had a major role in the revision of the original Wechsler Intelligence Scale for Children – published an influential book (Kaufman, 1979) that proposed a more sophisticated method for analyzing and interpreting WISC-R data. Kaufman's *intelligent testing* system was grounded on cognitive theories as well as factor analytic research. It started with the assumption that the FSIQ is inadequate as an explanation of a child's intellectual functioning and it used the reliability indices as well as the variety of measures provided by the WISC-R to generate more informative interpretive hypotheses to be supported or discarded in light of information derived from the test battery and from additional sources of data about the child.

The ideas that had been percolating for some time concerning the limitations of the traditional scales, as well as the possibility of developing intelligence tests that would reflect advances in theories of cognitive trait organization and that would apply the information collected in over six decades of factor analytic research on measures of cognitive abilities, gave impetus to the development of new and improved tests of intelligence.[5] In fact, some of these advances even began to be applied to the SB and the Wechsler scales with each successive revision. For example, the SB Fourth Edition (Thorndike, Hagen, & Sattler, 1986) used a model of cognitive abilities that incorporated the theory of fluid (*Gf*) and crystallized (*Gc*) intelligence (Horn & Cattell, 1966) as the middle level of a hierarchy with the *g* factor above it

5 It should be noted that group tests of abilities had been applying factor analytic findings in their development well before the 1970s.

and with four group factors – namely, verbal, quantitative, and abstract-visual reasoning as well as short-term memory – below it.[6] Similarly, after the death of David Wechsler in 1981, the scales that still bear his name started to explicitly incorporate a multifactor structure for grouping subtests in order to devise interpretive strategies rooted more firmly on an empirically defensible basis. The Wechsler scales published after 1990 have added new subtests as needed to shore up and clarify the factorial structure of the scales (see, e.g., Wechsler, 1991, 1997, 2003, and 2008). Thus, besides the Full Scale IQ, the other four major scores derived from the WISC-IV and the WAIS-IV, namely the Verbal Comprehension, Perceptual Reasoning, Working Memory, and Processing Speed composites, are based on groupings of subtest scores arrived at through factor analyses.

In addition to the structural revisions made by the traditional intelligence test batteries, a number of completely new instruments – with new scales and novel types of items – have also been appearing in the past few decades. Most of these make use to some extent or another of what has come to be known as the Cattell-Horn-Carroll (CHC) model of cognitive abilities. This model epitomizes the psychometric approach to intelligence pioneered by Spearman (1904, 1927) and pursued by many other investigators specializing in factor analysis of cognitive test data and in theories of cognitive trait organization. It consists of a hierarchical three-stratum arrangement devised by Carroll (1993) that serves to organize the massive amount of factor analytic research on human cognitive abilities accumulated over six or seven decades. The full model includes about 70 narrow abilities in the first or lowest stratum, approximately eight broad factors – including fluid and crystallized intelligence– in the second or middle stratum, and the general (*g*) intelligence factor in the third or highest stratum.

6 The Stanford-Binet 5th edition (Roid, 2003) uses a modified five-factor hierarchical model.

The Woodcock-Johnson III Test of Cognitive Abilities (WJ III; Woodcock, McGrew, & Mather, 2001), which is the current version of a test battery originally published in 1978, is one of the tests that has used the CHC model of cognitive abilities most extensively in its design, incorporating as it does seven of the CHC broad factors and over 20 of the narrow abilities in that model. Two other recent test batteries that use some aspects of the CHC model for their interpretive schemes are the Reynolds Intellectual Assessment Scales (RIAS; Reynolds & Kamphaus, 2003) and the second edition of the Differential Ability Scales (DAS-II; Elliott, 2007). In addition, the theory and research behind the CHC model, along with the intelligent testing method pioneered by Kaufman (1979, 1994), have been used to develop the *cross battery assessment* approach (XBA; Flanagan & McGrew, 1997; Flanagan, Ortiz, & Alfonso, 2007). This approach, as the name implies, offers guidance on how to design cognitive assessments using one of the comprehensive intelligence test batteries and supplementing it with additional tests from another intelligence or achievement battery, as may be required in light of the unique referral question to be addressed. Kaufman's intelligent testing provides an ideal basis for the utilization of the CHC. His method is geared toward understanding an examinee's pattern of cognitive strengths and weakness through the application of clinical and psychometric methods in a flexible and individualized fashion. The cross-battery approach is especially geared toward the evaluation of learning disabilities and toward the assessment of individuals from culturally or linguistically diverse backgrounds.

Developers of the new generation of intelligence tests have also employed the functional theory of brain organization developed by A. R. Luria and mentioned previously in connection with the K-ABC. This theory makes a distinction among functional units of the brain devoted primarily to attention, to planning, and to the successive and simultaneous processing of information.

Table 2.1. Major Examples of Current Intelligence Tests

Test Title and Acronym	*Author(s) and Date of Publication*	*Primary Theoretical/Empirical Rationale*
Cognitive Assessment System (CAS)	J. A. Naglieri & J. P. Das (1997)	PASS theory of cognitive functioning: Planning, Attention, Simultaneous, & Sequential Processing (Das, Naglieri, & Kirby, 1994)
Differential Ability Scales-Second Edition (DAS-II)	C. D. Elliott (2007)	Cattell-Horn-Carroll (CHC) model – Stratum II: Broad abilities (Carroll, 1993)
Kaufman Adolescent and Adult Intelligence Test (KAIT)	A. S. Kaufman & N. L. Kaufman (1993)	Horn and Cattell's (1966) model of Fluid (Gf) and Crystallized (Gc) intelligence & Luria's (1973, 1980) neuropsychological theory
Kaufman Assessment Battery for Children-Second Edition (KABC-II)	A. S. Kaufman & N. L. Kaufman (2004)	Luria's (1973, 1980) neuropsychological theory & Cattell-Horn-Carroll (CHC) model (Carroll, 1993)
Reynolds Intellectual Assessment Scales (RIAS)	C. R. Reynolds & R. W. Kamphaus (2003)	Cattell-Horn-Carroll (CHC) model – Stratum III: *g* & Stratum II: Broad abilities (Carroll, 1993)
Stanford-Binet Intelligence Scales-Fifth Edition (SB5)	G. H. Roid (2003)	Cattell-Horn-Carroll (CHC) model (Carroll, 1993) and factor analyses
Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV), Wechsler Intelligence Scale for Children-Fourth Edition (WISC-IV)	D. Wechsler (2008, 2003)	Factor analytically derived composites: Verbal Comprehension, Perceptual Reasoning, Working Memory, & Processing Speed
Woodcock-Johnson III Test of Cognitive Abilities (WJ III)	R. W. Woodcock, K. S. McGrew, & N. Mather (2001)	Cattell-Horn-Carroll (CHC) model-Stratum III, II, & I: *g* plus broad and narrow abilities (Carroll, 1993)

Successive processing involves serial or temporal sequencing of information whereas simultaneous processing involves synthesizing or organizing material as a whole and at once. As elaborated by J. P. Das and others (Das, Naglieri, & Kirby, 1994), Luria's conceptualizations were the foundation of the PASS theory of intelligence used as the primary basis for the development of the Cognitive Assessment System (CAS), an intelligence test battery authored by Das and Naglieri (1997). Alan and Nadine Kaufman, meanwhile, have also continued to use aspects of Luria's theory and of the Horn-Cattell model of *Gf* and *Gc* in developing the Kaufman Adolescent and Adult Intelligence Test (KAIT; Kaufman & Kaufman, 1993) and the second edition of the Kaufman Assessment Battery for Children (KABC-II; Kaufman & Kaufman, 2004). Table 2.1 lists the major examples of current intelligence test batteries, along with their authors and the theoretical or empirical rationale on which they are based.

Do Intelligence Tests Have a Future?

Here the short answer is, most likely, yes. As far as group tests of intelligence and

general aptitude are concerned, most of those listed in TIP can produce good estimates of general intellectual ability or *g*, provided their content is appropriate for the age, culture, educational background, and any special characteristics or disabilities of the examinee. They can also produce such estimates at low cost and without the need of extensive apparatus. With regard to the individually administered comprehensive intelligence test batteries that have been discussed here, the situation is somewhat different. To be sure, most of them can also provide good estimates of general intellectual ability and fulfill the original purpose for which the Binet and the Wechsler scales were developed. If that were all they could do, however, their cost and the extensive training required to properly administer them, score them, and interpret their results would not be justified.

The reason that individual intelligence tests are likely to endure is tied to their versatility and clinical usefulness. They essentially provide a standardized and structured interview script that the well-trained user can employ for gathering a broad sample of behavioral data relevant to cognitive functioning while observing stylistic variations that can also reveal clinically significant personality data. In the survey published by Camara et al. (2000), for example, out of the top 20 most frequently used tests, the WAIS-R was ranked in first place by clinical psychologists and in second place by neuropsychologists.[7] Not only have the traditional scales evolved and been improved with regard to their composition, psychometric properties, and normative bases, but a number of new ones have been published which expand the range of cognitive tasks that can be sampled and the array of empirical and theoretical evidence that can be adduced to support their validity. Thus, the utility of the tests for the assessment of adaptive/functional behavior, intellectual development, learning difficulties, neuropsychological and psychiatric problems, as well as for rehabilitation or remedial planning, has been greatly increased. Already, the procedures of some intelligence test batteries, notably the WISC-IV Integrated (Kaplan et al., 2004), have been modified so as to take advantage of the one-on-one administration mode to gather additional dynamic information on examinees' problem-solving processes and to contribute more directly to remediation planning. Furthermore, as Goldstein (2008) points out, recent advances in neuroimaging, such as the functional MRI, offer exciting possibilities for applying the more sophisticated and well-validated tasks of current tests to neurodiagnosis and to extending knowledge of brain-behavior relationships.

In a sense, nearly all of human behavior involves cognitive abilities as these encompass processes that include attention, perception, comprehension, judgment, decision making, reasoning, intuition, and memory, among others. Not all of these are tapped by intelligence tests (see, e.g., Stanovich, 2009). Nevertheless, the fact that the term *cognitive abilities* is increasingly used instead of intelligence – even in the titles of tests that might have been called "intelligence" tests in another era – is helpful because cognitive processes are more easily defined, grasped, and assessed and are not as emotionally laden as "intelligence" is. When the cognitive abilities tapped by intelligence tests are used in performing mental tasks or in problem solving, it is reasonable to assume that the one who is performing those tasks or solving those problems is displaying intelligent behavior. However, it also seems clear that not all intelligent behavior is simply a function of the cognitive abilities measured by the tests. What the tests do not measure, namely, characteristics such as motivation, flexibility, leadership ability, persistence, conscientiousness, and creativity, are as important as – or even more so than – the cognitive abilities the tests *do* measure in allowing individuals to behave intelligently and to cope with the challenges that life presents.

7 The MMPI, which was reported in the survey as the most frequently used instrument for personality assessment, was ranked in first place by neuropsychologists and in second place by clinical psychologists.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Psychological Association. (1974). *Standards for educational and psychological tests*. Washington, DC: Author.

Anastasi, A. (1986). Intelligence as a quality of behavior. In R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence? Contemporary viewpoints on its nature and definitions* (pp. 19–21). Norwood, NJ: Ablex.

Binet, A., & Simon, Th. (1904). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'Année Psychologique*, **11**, 191–244. Retrieved from http://www.persee.fr/web/revues/home/prescript/issue/psy_0003-5033_1904_num_11_1.

Boring, E. G. (1923, June 6). Intelligence as the tests test it. *New Republic*, **35**, 35–37.

Camara, W. J., Nathan, J. S., & Puente, A. E. (2000). Psychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice*, **31**, 141–154.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York, NY: Cambridge University Press.

Das, J. P., Naglieri, J. A., & Kirby, J. R. (1994). *Assessment of cognitive processes: The PASS theory of intelligence*. Boston, MA: Allyn & Bacon.

Elliott, C. D. (2007). *DAS-II administration and scoring manual*. San Antonio, TX: PsychCorp.

Fancher, R. E. (1985). *The intelligence men: Makers of the IQ controversy*. New York, NY: W.W. Norton.

Flanagan, D. P., & Harrison, P. L. (Eds.). (2005). *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed.). New York, NY: Guilford Press.

Flanagan, D. P., & McGrew, K. S. (1997). A cross-battery approach to assessing and interpreting cognitive abilities: Narrowing the gap between practice and cognitive science. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 314–325). New York, NY: Guilford Press.

Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2007). *Essentials of cross-battery assessment* (2nd ed.). Hoboken, NJ: Wiley.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin*, **95**, 29–51.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin*, **101**, 171–191.

Goldstein, G. (2008). Intellectual assessment. In M. Hersen & A. M. Gross (Eds.), *Handbook of clinical psychology* (Vol. 1, pp. 395–421). Hoboken, NJ: Wiley.

Gottfredson, L. S. (2009). Logical fallacies used to dismiss the evidence on intelligence testing. In R. P. Phelps (Ed.), *Correcting fallacies about educational and psychological testing* (pp. 11–65). Washington, DC: American Psychological Association.

Gould, S. J. (1996). *The mismeasure of man* (Rev. ed.). New York, NY: W. W. Norton.

Homack, S. R., & Reynolds, C. R. (2007). *Essentials of assessment with brief intelligence tests*. Hoboken, NJ: Wiley.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized intelligence. *Journal of Educational Psychology*, **57**, 253–270.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.

Kaplan, E., Fein, D., Kramer, J., Morris, R., Delis, D., & Maerlender, A. (2004). *WISC-IV Integrated: Technical and interpretive manual*. San Antonio, TX: PsychCorp.

Kaufman, A. S. (1979). *Intelligent testing with the WISC-R*. New York, NY: Wiley.

Kaufman, A. S. (1994). *Intelligent testing with the WISC-III*. New York, NY: Wiley.

Kaufman, A. S., & Kaufman, N. L. (1983). *Kaufman Assessment Battery for Children: Interpretive manual*. Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Kaufman, N. L. (1993). *Manual for the Kaufman Adolescent & Adult Intelligence Test (KAIT)*. Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Kaufman, N. L. (2004). *Manual for the Kaufman Assessment Battery for Children – Second Edition (KABC-II): Comprehensive Form*. Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Lichtenberger, E. O. (2006). *Assessing adolescent and adult intelligence* (3rd ed.). Hoboken, NJ: Wiley.

Kirk, R. E. (1999). *Statistics: An introduction* (4th ed.). Fort Worth, TX: Harcourt Brace.

Luria, A. R. (1973). *The working brain: An introduction to neuropsychology*. New York: Basic Books.

Luria, A. R. (1980). *Higher cortical functions in man* (2nd ed.). New York, NY: Basic Books.

McDermott, P. A., Fantuzzo, J. W., & Glutting, J. J. (1990). Just say no to subtest analysis: A critique of Wechsler theory and practice. *Journal of Psychoeducational Assessment*, *8*, 290–302.

Murdoch, S. (2007). *IQ: A smart history of a failed idea*. Hoboken, NJ: Wiley.

Murphy, L. L., Spies, R. A., & Plake, B. S. (Eds.). (2006). *Tests in Print VII*. Lincoln, NE: Buros Institute of Mental Measurements.

Naglieri, J. A., & Das, J. P. (1997). *Das-Naglieri Cognitive Assessment System*. Chicago, IL: Riverside.

Neisser, U. (Ed.). (1998). *The rising curve: Long-term gains in IQ and related measures*. Washington, DC: American Psychological Association.

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, **51**, 77–101.

Rapaport, D., Gill, M., & Schafer, R. (1945). *Diagnostic psychological testing* (Vol. 1). Chicago, IL: Year Book.

Rapaport, D., Gill, M., & Schafer, R. (1946). *Diagnostic psychological testing: The theory, statistical evaluation, and diagnostic application of a battery of tests* (Vol. 2). Chicago, IL: Year Book.

Reynolds, C. R., & Kamphaus, R. W. (2003). *Reynolds Intellectual Assessment Scales*. Lutz, FL: Psychological Assessment Resources.

Robertson, G. J. (1972). Development of the first group mental ability test. In G. H. Bracht, K. D. Hopkins, & J. C. Stanley (Eds.), *Perspectives in educational and psychological measurement* (pp. 183–190). Englewood Cliffs, NJ: Prentice-Hall.

Roid, G. H. (2003). *Stanford-Binet Intelligence Scales, Fifth Edition: Technical manual*. Itasca, IL: Riverside.

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures*. Retrieved from http://www.siop.org/_Principles/principles.pdf.

Sokal, M. M. (Ed.). (1987). *Psychological testing and American society: 1890–1930*. New Brunswick, NJ: Rutgers University Press.

Spearman, C. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology*, **15**, 201–293.

Spearman, C. (1927). *The abilities of man*. New York, NY: Macmillan.

Stanovich, K. E. (2009). *What intelligence tests miss: The psychology of rational thought*. New Haven, CT: Yale University Press.

Sternberg, R. J., & Detterman, D. K. (Eds.). (1986). *What is intelligence?* Norwood, NJ: Ablex.

Sundet, J. M., Barlaug, D. G., & Torjussen, T. M. (2004). The end of the Flynn effect? A study of secular trends in mean intelligence scores of Norwegian conscripts during half a century. *Intelligence*, **32**, 349–362.

Teasdale, T. W., & Owen, D. R. (2005). A long-term rise and recent decline in intelligence test performance: The Flynn effect in reverse. *Personality and Individual Differences*, **39**, 837–843.

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986). *The Stanford-Binet Intelligence Scale: Fourth Edition, Guide for administering and scoring*. Chicago, IL: Riverside.

Thurstone, L. L. (1934). The vectors of mind. *Psychological Review*, **41**, 1–32.

Urbina, S. (1993). Intelligence: Definition and theoretical models. In F. N. Magill (Ed.), *Survey of social science: Psychology*. Pasadena, CA: Salem Press.

Wechsler, D. (1958). *The measurement and appraisal of adult intelligence* (4th ed.). Baltimore, MD: Williams & Wilkins.

Wechsler, D. (1975). Intelligence defined and undefined: A relativistic appraisal. *American Psychologist*, **30**, 135–139.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children – Third Edition*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale – Third Edition*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (2002). *Wechsler Preschool and Primary Scale of Intelligence – Third Edition*. San Antonio, TX: Harcourt Assessment.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children – Fourth Edition*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale – Fourth Edition*. San Antonio, TX: Pearson.

Wissler, C. (1901). The correlation of mental and physical tests. *Psychological Monographs*, 3(6), 1–62.

Wolf, T. H. (1973). *Alfred Binet*. Chicago, IL: University of Chicago Press.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III*. Itasca, IL: Riverside.

Zenderland, L. (1998). *Measuring minds: Henry Herbert Goddard and the origins of American intelligence testing*. New York, NY: Cambridge University Press.