Canadian Psychology
2009, Vol. 50, No. 3, 183–195

0708-5591/09/$12.00 DOI: 10.1037/a0016641

# Intelligence: Foundations and Issues in Assessment

Linda Gottfredson
University of Delaware

Donald H. Saklofske
University of Calgary

There is no more central topic in psychology than intelligence and intelligence testing. With a history as long as psychology itself, intelligence is the most studied and likely the best understood construct in psychology, albeit still with many "unknowns." The psychometric sophistication employed in creating intelligence tests is at the highest level. The authors provide an overview of the history, theory, and assessment of intelligence. Five questions are proposed and discussed that focus on key areas of confusion or misunderstanding associated with the measurement and assessment of intelligence.

*Keywords:* intelligence, intelligence testing, intelligence tests

To understand how we are like all others, some others, and no others, to paraphrase Kluckhohn and Murray (1948), has led to the search for those key individual differences factors that can be operationally defined and measured. Individual differences characteristics attributed as the causes or underlying basis of human behaviour include intelligence, personality, and conative factors. Here, the view is that school achievement, job success, or longevity is, in turn, related to these latent traits. Thusly, school success is causally related to, or at least correlated, for example, with fluid intelligence, memory, and processing speed, on the one hand, and personality factors such openness to experience and conative factors such as motivation, on the other (Hilgard, 1980). And, if we know how these factors are distributed in the population, following, say, normal curve expectations, then future behaviours may be predicted using reliable and valid measures of these latent traits and, of course, the criterion measure as well.

## A Historical Note on Intelligence Tests

Defining and measuring intelligence predates scientific psychology founded in the mid-19th century. Psychology provided the needed forum for the study of intelligence as a key individual differences factor. Although the elementary "brass instruments" tests created and used by Galton and James McKeen Cattell raised interest in the measurement of intelligence, it was the practical success of the Binet–Simon tests in France at the beginning of the 20th century and their adoption in the United States that propelled the study and measurement of intelligence into its current central position in both the discipline and practise of psychology (see Boake, 2002; Tulsky et al., 2003). Whipple stated in the preface to his 1914 book *Manual of Mental and Physical Tests*,

> One need not be a close observer to perceive how markedly the interest in mental tests has developed during the past few years. Not very long ago attention to tests was largely restricted to a few laboratory psychologists; now tests have become objects of the attention for many workers whose primary interest is in education, social service, medicine, industrial management and many other fields in which applied psychology promises valuable returns. (Whipple, 1914, p. v)

The Army Alpha and Beta tests used extensively for screening U.S. military recruits during World War I further demonstrated the practical relevance of assessing intelligence quickly and inexpensively in diverse adult groups. Companies were founded to create these tests (e.g., the Psychological Corporation, now Pearson Assessment, formed by James McKeen Cattell et al.), which were then catalogued and critically reviewed in the *Mental Measurements Yearbook*, first published in 1938 by Buros.

By the 1930s, the diversity in the content and method of intelligence assessment included specific tests such as the Porteous Maze, Kohs Block Design, Goodenough Draw-a-Man, and Raven Progressive Matrices, and more general mental ability tests such as the Munroe–Buckingham General Intelligence Scale and Stutsman Merrill–Palmer Scale of Mental Tests. Very young children could be assessed using the Cattell Infant Intelligence Scale. Intelligence tests were used in schools for assessing underachievement, mental retardation, giftedness, and the abilities of children presenting with conditions that might interfere with learning (e.g., deafness, visual impairments). The confidence placed in these tests is exemplified by the Psychological Corporation's description of the Wechsler Bellevue Intelligence Scale as an "individual examination including 10 subtests at any level . . . translated into standard score units . . . converted into IQ equivalents by reference to a table . . . well suited for classification . . . norms for 7–70 years." *And the price was only $12.50!!* (see Tulsky et al., 2003).

## Concurrent Development of Theories of Intelligence and Intelligence Tests

The Galton-type measures were not grounded in theory to guide their development, interpretation, and integration into a fuller description of human behaviour, and their "narrowness" did not allow for, say, predicting school or occupational success. The early theoretical underpinnings of intelligence are found in Charles Spearman's (1904) two-factor intelligence model describing specific or "*s*" factors (akin to primary factors) and a second-order general factor or "*g*." In

Linda Gottfredson, School of Education, University of Delaware; Donald H. Saklofske, Division of Applied Psychology, University of Calgary.

Correspondence concerning this article should be addressed to, E-mail: lindagottfredson@gmail.com

contrast, E. L. Thorndike (1924) viewed intelligence as several unique factors, whereas Louis Thurstone (1938) proposed 7 uncorrelated factors, each of which could be measured and described separately using the Primary Mental Abilities Test. J. P. Guilford (1967) hypothesised three broad intelligence factors (i.e., operations, content, and products) defined by some 120 or more specific factors, each requiring a specific test. The seminal contributions of John Horn and R. B. Cattell (1966) describing crystallized (*Gc*) and fluid (*Gf*) intelligence were seen by some as paralleling the Verbal (VIQ) and Performance IQ (PIQ) scores of the Wechsler tests. More recently, multifaceted descriptions of intelligence include Sternberg's (1997) triarchic theory, Gardner's (1983, see also Gardner & Karnbaber & Wike, 1996) multiple intelligences, and the Naglieri–Das (Naglieri, 2009) PASS model (i.e., planning, attention, simultaneous, and successive/sequential processing).

John Carroll's (1993) review and analysis of the large intelligence database resulted in a three-stratum model of human intelligence and cognitive abilities. This model is regarded by many as the best representation of the "structure of human cognitive abilities" because of the strength of its empirical foundation. Combining the work of Cattell and Horn with Carroll's description of intelligence, McGrew and Flanagan (1998) proposed an integrated model referred to as CHC theory that serves as the foundation for the Woodcock–Johnson III Tests of Cognitive Abilities (WJ-III; Woodcock, McGrew, & Mather, 2001) and has been applied to alternative interpretations of the Wechsler Intelligence Scale for Children—Fourth Edition (WISC–IV; Wechsler, 2003) and the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS–IV; Wechsler, 2008).

These hierarchical models and the *Gf/Gc* perspective along with the CHC theory are variously reflected in current tests. All Wechsler tests include a second-order *g* or full-scale IQ (FSIQ) score following the Spearman tradition. VIQ and PIQ have a resemblance to Vernon's (1950) verbal–educational (V:Ed) and spatial–mechanical (K:M), and the four factors of the WAIS–IV and WISC–IV certainly tap several of Carroll's Stratum II factors, including *Gf* and *Gc*, while the various subtests of these scales represent at least some of the narrow CHC factors akin to Vernon's minor group factors or even very specific factors.

The lack of agreement on a theory and structure of intelligence is the major reason for so many intelligence tests. Comprehensive tests such as the WISC–IV, WAIS–IV, WJ-III, and Stanford–Binet Intelligence Scales—Fifth Edition (Roid, 2003) share a number of similar or overlapping components; all yield a second-order general factor that is at least moderately correlated across these different tests. Grounded in alternative perspectives, the PASS model, reflecting brain–behaviour relationships initially developed by Luria, is tapped by the Cognitive Assessment System (Naglieri, 2009; Naglieri & Das, 1997). Operating from a neo-Piagetian perspective and drawing from Vygotsky's description of the "zone of proximal development" is the "dynamic assessment" approach (e.g., Tzuriel, 2001). Other large-scale models such as those developed by Sternberg and Gardner have not led to actual tests or test batteries in the more traditional sense but do have heuristic value.

A number of single factor and "brief " intelligence tests assess specific abilities (e.g., Raven's Progressive Matrices) or focus on particular clients (e.g., visually impaired). The Wechsler Abbreviated Scale of Intelligence (Wechsler, 1999) employs two or four subtests to yield VIQ, PIQ, and FSIQ scores. The Wechsler Nonverbal Scale of Ability (Wechsler & Naglieri, 2006) and Universal Nonverbal Intelligence Test (Bracken & McCallum, 1998) have reduced the need for language by both the examiner and client.

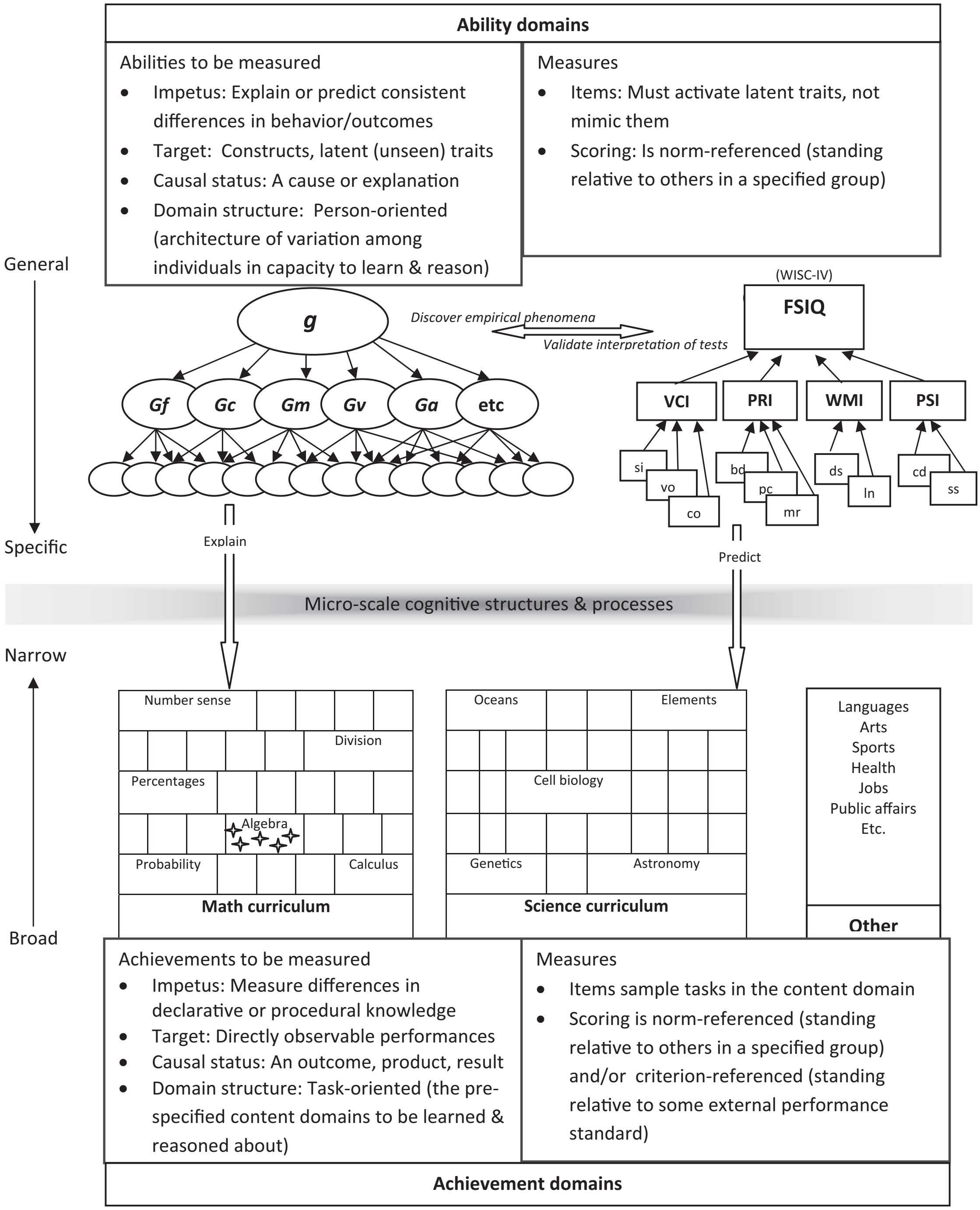
## "Under Construction": In-Development Models of Intelligence

The study of intelligence involves almost all areas of psychology and other sciences such as neurobiology and behaviour genetics. This has expanded our view of what intelligence is, how it develops, and the many factors that influence it, such as aging (see Deary, Whalley, & Starr, 2009; Kaufman, 2008; Lee, Gorsuch, Saklofske, & Patterson, 2008). It is well established that intelligence is a product of both hereditary and environmental factors and their interaction. But the complexity of intelligence and how we measure it and interpret these measures go much further. Intelligence does not exist or affect human behaviour in isolation. Hans Eysenck (1997) contended that intelligence and personality are orthogonal constructs, but he also was quick to point out that there is a big difference between a bright and less intelligent extravert! Several books have examined the interface between intelligence and personality (e.g., Saklofske & Zeidner, 1995), and Collins and Messick (2001) further argued that conative factors such as motivation interact with intelligence. More recently, Ackerman (Ackerman & Beier, 2003a,b; Ackerman & Kanfer, 2004) has described the interaction amongst cognitive, affective, and conative variables and career choice process, and trait determinants of expertise. Interactive models require a mixed methods approach to the assessment (formal and informal tests, interviews, observation, case history, etc.) of multiple individual differences variables in order to yield a comprehensive and integrated description of the client.

We now have a much better understanding of intelligence and its relationship to other related factors such as memory, achievement, and executive functioning. Co-norming studies completed during the standardisation of the third edition of the WAIS (WAIS–III) and Wechsler Memory Scale—Third Edition (WMS–III) later resulted in a proposed six-factor model of memory and intelligence (Tulsky et al., 2003). During the standardisation of the WAIS–IV, data were collected on other measures such as the fourth edition of the WMS, Wechsler Individual Achievement Test—Second Edition (WIAT–II), Delis–Kaplan Executive Function System (D-KEFS), California Verbal Learning Test—Second Edition (CVLT–II), and Repeatable Battery for the Assessment of Neuropsychological Status (RBANS). More to the point, these standardisation studies provide a stronger empirical foundation for the clinical assessment of clients complemented by the plethora of research publications that further enhance the clinical utility of these tests. For example, methods have been developed for estimating premorbid intelligence on the basis of WISC–IV and WAIS–III scores and demographic variables (e.g., Schoenberg, Lange, & Saklofske, 2007; Schoenberg, Lange, Saklofske, Suarez, & Brickell, 2008). Psychologists can now examine multiple cognitive and other factors where the relationship between the measures is known, as are their respective reliabilities; given the same or overlapping normative samples, it is possible to more precisely predict and describe or diagnose behaviours of interest.

## Assessing Intelligence: The Good, Bad and . . .!

Intelligence testing has been amongst the most controversial topics in psychology and other professional arenas such as educa-



**Ability domains**

Abilities to be measured

- Impetus: Explain or predict consistent differences in behavior/outcomes
- Target: Constructs, latent (unseen) traits
- Causal status: A cause or explanation
- Domain structure: Person-oriented (architecture of variation among individuals in capacity to learn & reason)

Measures

- Items: Must activate latent traits, not mimic them
- Scoring: Is norm-referenced (standing relative to others in a specified group)

Achievements to be measured

- Impetus: Measure differences in declarative or procedural knowledge
- Target: Directly observable performances
- Causal status: An outcome, product, result
- Domain structure: Task-oriented (the pre-specified content domains to be learned & reasoned about)

Measures

- Items sample tasks in the content domain
- Scoring is norm-referenced (standing relative to others in a specified group) and/or criterion-referenced (standing relative to some external performance standard)

**Achievement domains**

*Figure 1.* Psychometric distinction between ability and achievement domains, constructs, and tests.

tion as well as amongst the general public. The construct of intelligence was and still is not universally embraced by all psychologists. Radical behaviourism, social constructivist, interpretist, and other movements within and outside psychology see little value in including intelligence as a basis for describing the human condition. The use of intelligence tests for classifying and placing children into special classes has been challenged in the U.S. courts during the past 30 years.

In the 1990s, key publications such as *The Bell Curve* (Herrnstein & Murray, 1994) presented intelligence as one of the most important predictors of a wide range of human behaviour. Further revisited were the issues of race differences in intelligence and the role of heredity, genetics, and environmental factors. The strong and outspoken reaction to this book resulted in Gottfredson's 1994 *Wall Street Journal* editorial (reprinted in *Intelligence*, 1997) entitled "Mainstream Science on Intelligence" that was endorsed by

51 other leading scientists supporting the content and conclusions presented by Herrnstein and Murray.

Concurrently, an "expert" task force was assembled by the Board of Scientific Affairs, American Psychological Association, to write a position paper in response to the dogma and non–evidence-based opinion regarding intelligence but also to put an informed lens on the vast psychological literature, including *The Bell Curve*. "Intelligence: Knowns and Unknowns," first made public in 1995, was published the following year in the *American Psychologist* (Neisser et al., 1996). This was followed by another American Association of Psychologists task force (Psychological Assessment Work Group) examining psychological testing and assessment. Here, Meyer et al. (2001) concluded,

> (a) Psychological test validity is strong and compelling, (b) psychological test validity is comparable to medical test validity, (c) distinct assessment methods provide unique sources of information, and (d) clinicians who rely exclusively on interviews are prone to incomplete understandings . . . the authors suggest that a multimethod assessment battery provides a structured means for skilled clinicians to maximize the validity of individualized assessments. Future investigations should move beyond an examination of test scales to focus more on the role of psychologists who use tests as helpful tools to furnish patients and referral sources with professional consultation. (p. 128)

Although the focus of the Meyer et al. study was on all psychological tests and testing, there is also very good evidence in support of the psychometric integrity and clinical applications of current intelligence tests. Comparing the 1939 WB with the 1997 WAIS–III highlights the many significant changes in the test items, subtests, IQ/index scores, psychometric properties, factor structure, standardisation and norming methods, scoring criteria, and empirically supported interpretation strategies (Tulsky et al., 2003). The newly published WAIS–IV further expanded the contemporary theoretical foundation, developmental appropriateness, user friendliness, and clinical utility as well as providing current norms. An examination of the WAIS–IV psychometric properties alone is reassuring. The reliability of the FSIQ is .98 (with an average *SEM* of 2.16) and .90 to .96 for the four index scores. Confirmatory factor analysis results support a four-factor structure together with a second-order *g* factor. Furthermore, this factor structure for the WAIS–III has been replicated in other countries such as Canada (Bowden, Lange, Weiss, & Saklofske, 2008). Even more compelling if one looks to the Wechsler children's scales is the cross-cultural robustness of the WISC–III on the basis of data from North American, European, and Asian countries (Georgas, Weiss, von de Vijver, & Saklofske, 2003).

A major criticism of earlier intelligence testing has been that other than for intellectual classification purposes, these test scores yield little of value for diagnosis or psychological prescription. This has changed markedly in recent years as intelligence has become more integrated into describing the "whole person" in relation to academic placement, job selection, or potential for recovery from brain injury, for example. The current emphasis on positive psychology has also highlighted the importance of intelligence as a key resiliency factor and in building "capacity" and psychological well-being. Prescriptive recommendations have been developed for several of the more often used current intelligence tests such as the WISC–IV (Prifitera, Saklofske, & Weiss, 2008).

Anastasi and Urbina (1997) concisely stated that we are all of equal value as human beings while at the same time differing in many ways. Added to this is the commonly heard adage in the measurement and assessment literature that tests are neutral; it is what we do with them that makes them good or bad, useful or useless. Thusly, the picture is not complete until we also determine how we will use this information and its impact on the individual and society. The *Standards for Educational and Psychological Testing* are again to be revised (see APA Press Release, September 24, 2008), and organisations such as the International Test Commission strive to promote effective testing and assessment practises and advocate for the proper development, evaluation, and uses of educational and psychological instruments. Best practises in assessment are described in such publications as the *Principles for Fair Student Assessment Practises for Education in Canada* (1993), and the ethics codes adhered to by psychologists guide all areas of test use, assessment, and diagnosis (e.g., Canadian Psychological Association, 2000; Universal Declaration of Ethical Principles for Psychologists, 2008).

## Five Major Trends and Controversies in Intelligence Assessment

The following five questions seem particularly useful in capturing the guiding concerns over which protagonists have sparred and shaped the course of intellectual assessment during the past half century. In addressing the questions, we draw on two key distinctions illustrated in Figure 1: (a) abilities versus achievements (potentials vs. accomplishments), and (b) constructs versus measures (the phenomena that testers aim to capture vs. the tools they use to do so). Many controversies arise from not distinguishing between them.

The ability–achievement distinction is illustrated vertically in Figure 1. In the context of intellectual assessment, abilities are conceived as causes and achievements as outcomes. Achievements are accomplishments that can be directly observed. Abilities are not directly observable, but are *latent traits* we infer from regularities in behaviour that we notice across time, place, and circumstance. To illustrate, whereas accomplishments such as proficient spelling, grammar, and word choice can be directly observed (and measured), differences in verbal ability—a general potential for verbal achievements—must be inferred from various signs.

Achievements are, by definition, valued human products and performances that we want institutions to promote, such as reading and writing with proficiency. Achievements are thusly content specific and culture specific. Creating an achievement test requires carefully delineating the specific domain of content from which to sample test takers' knowledge, say, of algebra or French grammar. The achievement outcomes to be assessed are therefore enumerated a priori.

Unlike achievement domains, the empirical architecture of abilities—their organisation and relatedness—must be discovered. Abilities are latent traits, or psychological *constructs*—relatively stable differences amongst individuals in their propensities and capacities. Psychological constructs represent hypotheses about unobservable causal forces, or organising tendencies, in the human psyche. Psychologists posit a wide variety of them, such as intelligence, extraversion, and emotional stability, to explain why individuals are so consistently different across time and circumstance.

Latent traits are hypothesised causal forces whose influence can be mapped and described; we cannot create, nullify, or reconfigure their causal powers by acts of verbal definition or specification. In contrast, we may classify any observed behaviour we wish as an achievement, depending on which human performances we value.

The second distinction—construct versus measure (phenomenon vs. yardstick used to measure it)—is represented horizontally in Figure 1. Tests are standardised means of instigating samples of behaviour that will, if properly quantified and interpreted, reveal individual differences in the latent traits (abilities) or developed competencies (achievements) we wish to assess. When tests are used to measure achievement in some content domain, test items must systematically sample a carefully delineated domain of declarative or procedural knowledge. This analytical process of circumscribing and sampling a content domain is represented in Figure 1 for algebra by the stars within that segment of the much larger domain of math achievement. Test takers' responses on valid tests of achievement therefore look like, and indeed are, the thing itself—achievement in that specific content domain, or at least a representative sample of it. The validity of an achievement test rests on its content validity; test content must closely match task content within the achievement domain.

The situation is altogether different for tests of latent abilities like intelligence. Here, test items need not look like the hypothesised ability, but just set it in motion to produce observable, quantifiable responses. So, whereas achievement tests sample a domain of outcomes, ability tests must provoke the unseen phenomenon to reveal itself through its effects on behaviour. The test must set the hypothetical construct in motion to *cause* observable outcomes. The resulting test behaviour is not the causal force itself, but its product. Validating an ability test thusly requires showing that the test produces the patterns of effects across tasks and individuals that we would expect the hypothesised cause to create.

In Figure 1, latent constructs are represented in the upper left and measures of them are in the upper right. The difference between a construct and its measure is muddied when the terms *IQ* (a test score) and *intelligence* (a construct) are used interchangeably, as they commonly are. Test format and content need not mirror what we imagine the unobservable trait or causal force to "look like." That is why neither item content nor test format provides evidence for or against a test's validity for measuring the intended latent construct. As noted, test items need only activate the ability under controlled circumstances. Establishing construct validity requires evidence that the behaviour elicited by the test is consistent with propositions about the construct supposedly activated.

### *Question 1: What Is intelligence? Trend Away From Debating Definitions Toward Debating Discoveries*

A scientific answer to this question would describe an empirical phenomenon, not a test. It would therefore derive, not from a priori verbal definitions, but from extensive observation and experimentation. A full answer would cross disciplines and levels of analysis to explain the roots, meaning, and social consequences of human cognitive diversity. That is, explaining "what intelligence is" as a *construct* requires accumulating an expansive network of evidence that is consistent and coherent—consilient. IQ tests are merely tools we design to corral and examine the latent trait, whatever it turns out to be. But designing tests of unverified latent constructs necessarily begins with a belief that an important trait exists to be measured and a crude initial hypothesis about what it might be.

Binet's contribution to intellectual assessment was to conceptualize it in a manner that allowed him to measure it tolerably well. On the basis of his extensive observations of children, he posited that learning at grade level requires age-typical growth in a general reasoning capacity that is also manifest in everyday learning. If so, then bits of knowledge and skill that children acquire in daily life could be used to distinguish those who are proficient in picking them up from those less proficient. This strategy did, in fact, yield a test that identified children who were unlikely to succeed in the elementary curriculum without special assistance. The great practical utility of Binet and Simon's measuring device did not thereby validate Binet's notions of intelligence (the latent construct), but it did ignite an explosion of testing.

Variations on their test were enthusiastically adopted in ensuing decades by mass institutions seeking to select and place individuals in a more efficient and fair manner. Calculating a predictive validity is a straightforward statistical exercise, and for many practical purposes, it does not matter what a test actually measures or why it predicts some valued outcome. What matters to the institution is that the test does its assigned job, perhaps to select military recruits who can pass basic training or to place them in specialties for which they have the requisite aptitudes. In such cases, it is not necessary to validate, acknowledge, or even know the latent traits being measured. Hence, many de facto tests of intelligence go by other names (e.g., the Armed Forces Qualifying Test, which assesses "trainability") or just acronyms (the SAT, which assesses "college preparedness"). It might be useful to know what construct is being measured, but it can also be politically or economically expedient to avoid identifying or naming it.

Although construct validity is not required for many operational uses, lack of compelling evidence for it invites controversy when tests are perceived to have social consequences. For instance, sociologists in the 1970s often argued that IQ tests do not measure intellectual abilities at all; rather, they thought, intelligence tests predict school or job performance simply because they reflect a person's social class advantage. By their reasoning, intelligence is just a stand-in for social privilege and has no independent existence or value, except for what we arbitrarily assign to it. Earlier in the century (Spearman vs. Thurstone), and later as well (Jensen vs. Gardner and Sternberg), psychologists debated whether there exists any single, general intelligence or, instead, various semi-independent broad abilities. Stated another way, the question was whether there is any unitary construct out there to be measured, or whether "overall intelligence" is just the summation of independent or semi-independent abilities that we might choose to add, or not, to the IQ measurement pot.

All these questions are about latent constructs, not tests. Their answers have been provided primarily by empiricists interested in the origins, structure, distribution, and social implications of human intelligence (the construct), not in the pragmatics of tests and measurements (the yardsticks). A century of data collection with many tests, hundreds of researchers, and millions of test takers has revealed much about the phenomena that IQ test batteries capture.

It has provisionally settled the question of what intelligence is at the psychometric level, and started to answer questions about the developmental course, physiological correlates, genetic roots, and social consequences of human variation in intelligence. We now highlight three broad scientific conclusions about that variation: its psychometric structure, functional utility, and biological basis.

The first is that there are many abilities, but all are systematically interrelated with the others. The upper left portion of Figure 1 depicts an abbreviated version of Carroll's (1993) hierarchical model of intelligence. When tests of specific abilities are factor analysed, they yield a small number of common factors, generally corresponding to the so-called primary or group abilities such as verbal and spatial ability. These factors are themselves correlated, often strongly so, and thusly when factor analysed, they yield a yet more general, higher level common factor, called *g* (for the general mental ability factor), which is shown at the apex of the hierarchy. This hierarchical organisation of ability factors thereby integrates seemingly divergent perspectives on intelligence into a single framework. Moreover, the same model seems to fit all ages, sexes, races, and cultures yet examined (Jensen, 1998).

A particularly important point illustrated by the hierarchical model is that latent abilities are distinguished primarily by their breadth of application across content domains and only secondarily by content itself. This integrative model has settled the one-versus-many-intelligences question for most intelligence researchers, partly by showing that there emerges only a single highly general factor, *g*, at the highest level (Carroll's Stratum III). The *g* factor closely lines up empirically with full-scale IQ scores and conceptually with what most people think of as intelligence. Moreover, *g* is unitary at the psychometric level, that is, not a mixture of more specific abilities. Quite the reverse: It contributes the common core to all tested mental abilities. The narrowest abilities are the more factorially complex amalgams, as indicated by the downward-pointing arrows in the hierarchical model. (Note that the arrows go the opposite way for tests, because the more global scores are calculated by summing scores on the narrower tests. This scoring procedure is often misunderstood to represent how the latent traits themselves are constituted.)

Each of the broad ability factors at the next lower Stratum II level of generality, five of which are depicted in Figure 1, enhance performance in different content domains. The more cognitive of Gardner's "multiple intelligences" seem located in this stratum of generality (e.g., "visuospatial" ≈ spatial visualisation [Gv], "musical" ≈ auditory perception [Ga]), which illustrates that they are not nearly as independent as often presumed. The non-*g* components of broad abilities account for meaningful but relatively small amounts of variance in test scores.

As mentioned earlier, this hierarchical model is sometimes referred to as the Cattell–Horn–Carroll (CHC) model of intelligence because it accommodates the distinction between fluid and crystallized intelligence, introduced by Raymond Cattell and elaborated by John Horn. Fluid *g* represents raw information-processing power, which facilitates fast and accurate learning and reasoning in novel situations especially. Crystallized *g* represents primarily language-based capacities accrued from deploying one's fluid *g* in the past, that is, as a result of investing fluid *g* earlier in life. Cattell and Horn chose not to extract a higher level factor from their two highly correlated factors. Carroll locates both in Stratum II, where fluid *g* represents a general facility to reason well (including quantitatively) and crystallized *g* represents a general facility in verbal comprehension and communication (e.g., language, reading, listening). Fluid *g* is difficult to distinguish from the general factor *g*, and some research finds fluid *g* isomorphic with *g* itself.

For researchers, the label *intelligence* no longer has scientific meaning because it is attached to such diverse phenomena: often to *g* alone but sometimes to the more domain-specific factors at Stratum II (ignoring *g*), to the entire hierarchical structure of cognitive abilities, or even to that entirety plus a plethora of noncognitive forms of adaptive behaviour as well. Debates over how to define *intelligence* are now moot because the various empirical referents to which the term is commonly applied can be distinguished empirically and related within a common conceptual structure. This alone is a big advance in understanding human intelligence.

The second conclusion is that these measured abilities (latent constructs) are not psychometric chimera; they are not created by the measurement process but are important phenomena in the real world. They affect people's life chances regardless of whether we ever measure them. We focus here on *g* because, being the dominant factor, it is both the best understood and most consequential overall. As manifested in behaviour, it is a generalised capacity to learn, reason, and solve problems in virtually any content domain. Its utility increases, however, with a task's complexity, by which is meant the complexity of information processing it requires for good performance. Task complexity increases, for example, with greater amount and abstractness of information to integrate; more irrelevant information to ignore; a need to draw inferences or select appropriate procedures; and greater uncertainty, unpredictability, and fluidity of task, information, and circumstance. For example, *g* correlates somewhat with performance in all jobs but more strongly in more complex jobs (Schmidt & Hunter, 1998).

The utility of higher *g* also increases when performance outcomes are evaluated more objectively or depend more exclusively on the individual's own efforts and good judgement. For example, scores on *g*-loaded tests correlate more highly with objectively ascertained performance on a job than with supervisor ratings and more with on-the-job performance than earnings. Moreover, *g* accounts for the lion's share of prediction achieved by any broad battery of cognitive tests. Narrower abilities (e.g., spatial visualisation) contribute to prediction, but to a far lesser degree than does *g* and in a narrower range of activities (some sciences, crafts jobs, and graphic arts). Most life arenas require continual learning and reasoning, which may explain why *g* predicts such a wide range of life outcomes to at least some degree, from educational achievement to physical health. Individuals of lower IQ are at higher risk of school dropout, inadequate functional literacy, and adult poverty, but also accidental injury, preventable chronic illnesses, lower adherence to medical treatment, and premature death (Deary, in press; Gottfredson, 1997).

Third, cognitive diversity is a highly predictable, biological feature of all human populations (Bouchard, 1998; Jensen, 1998). To illustrate, all children follow much the same trajectory of cognitive development but at somewhat different rates, and therefore reach somewhat different levels by the time mental growth plateaus; the resulting interindividual differences in IQ within an age cohort stabilise by the time raw mental power (fluid *g*) peaks in early adulthood; those individual differences correlate at both the phenotypic and genetic levels with diverse structural and functional features of the brain (e.g., volume of the whole and individual parts, lower glucose metabolism, dendrite length, latency and amplitude of resting and evoked electrical potentials); the heritability of individual differences in cognitive ability increases with age, up to 80% amongst those surviving to late adulthood; the more general abilities are the more heritable and less trainable; cognitive faculties (learning, reasoning, and other aspects of fluid intelligence) weaken as the body ages physiologically; and selected interventions can improve or restore some cognitive functions, but the improvements tend not to generalise or be permanent, perhaps because they may not change *g* itself.

The raw scores on tests of fluid ability seem to rise and decline together, as a general package, and in tandem with physical maturation and decline across the life course. But this is not the case with crystallized *g*: It seems less vulnerable to physiological aging because it tracks the rise of fluid *g* but not its decline. Its usual maintenance into late adulthood makes sense if it represents the networks laid down in the brain of information and skills that were gained earlier through sustained learning and reasoning—investment—in broad content areas such as one's language and civilization. This now-automatized knowledge may decay or become harder to access, but it is not the machinery required to manipulate old and new information—that would be fluid *g*. The trajectories of an individual's fluid and crystallized *g* thusly diverge (within the individual) with advancing age, but individual differences in fluid and crystallized *g* still remain highly correlated (across individuals) at all ages.

To summarise, debates over the nature of intelligence have shifted away from debating arm-chair conceptions of intelligence to charting its empirical terrain. The accumulated research on individual differences in cognitive ability reveals that the seemingly myriad abilities we measure with mental tests are actually organised around a small number of dimensions at the psychometric level, that differences in cognitive ability have pervasive practical consequences in daily affairs, and that they are conditioned by the genetic diversity that is humankind's biological heritage. Not that all cognitive diversity is genetic in origin—most certainly not—but human genetic diversity does guarantee much phenotypic diversity in all populations. The job of intellectual assessment is first to capture that phenotypic diversity (or selected segments of it) and then to locate given individuals within it.

### *Question 2: What Do Intelligence Tests Actually Measure? Trend Away From Validating Yardsticks According to Intent, Appearance, and Similarity of Results to Validating the Conclusions Drawn From Them*

We turn now from interrogating latent constructs (upper left of Figure 1) to interrogating tests (upper right of Figure 1). The question here is whether the scores from tests adequately capture the reality in ability patterns that has been pieced together over the decades. Tests do not necessarily measure the constructs their developers intended or believe they do, and plenty of testing's critics have opined that IQ tests do not measure intelligence, at least as they construe it.

In the beginning, there was no hierarchical model against which to validate the construct validity of particular tests. In terms of Figure 1, no conception of intelligence had sufficient empirical evidence to lay claim to its upper left quadrant. It remained hotly contested ground for most of the century. Claims for the construct validity of the first intelligence tests were, of necessity, based mostly on the logic by which they were constructed, the tests' ability to discriminate amongst individuals perceived as bright versus dull, and their success in predicting the sorts of achievements that a general intelligence would be presumed to facilitate. Such limited evidence clearly does not rule out alternative hypotheses, so there followed much controversy over what intelligence tests actually measure. Test manuals continue to cite the high correlations of their battery's full-scale IQ scores with the IQs from other batteries, but such overlap does not by itself tell us what it is they are measuring in common.

We describe two once-plausible alternatives, disproved by the research just reviewed, that are still oft-invoked by testing's detractors (Gottfredson, 2009). We then turn to how construct-related research has increased our ability not just to rule out such alternatives, but also to evaluate, compare, and improve the construct validity of particular ability tests.

The once-plausible hypothesis conflates ability and achievement, and the second collapses the distinction between construct and measure. Both reflect the behaviourism of earlier eras in psychology, which eschewed anything not directly observable. The false ability-equals-achievement hypothesis may be paraphrased as follows: Tests of ability (upper right quadrant of Figure 1) actually gauge what an individual already knows or has accomplished in particular content domains (see lower half of Figure 1), not their standing on some hypothesised inner property or a latent construct (i.e., aptness in acquiring the necessary knowledge and skills; see upper left quadrant of Figure 1). A specific example would be the assertion that the Armed Forces Qualifying Test measures only one's history of exposure to instruction in school. This is essentially a claim that intelligence does not exist as a latent trait, that there are no individual differences in intellectual prowess to be measured.

The yardstick-equals-phenomenon hypothesis is the false assertion that, to be valid, an ability test must use items that look like or closely mimic the latent trait in question. By this logic, a test measures only that which it superficially resembles. Not surprisingly, observers who try to read the tea leaves of test content and format often interpret them differently: for instance, variously asserting that standardised tests obviously measure only the ability to take tests, or do academic-seeming work, or solve esoteric, prespecified problems with unambiguously right or wrong answers.

The discovery of the hierarchical organisation of latent abilities warns us against giving undue credibility to the name, appearance, or intent of any test. There are many more names attached to ability tests than there are latent ability factors, and manifest test content is a poor guide to the latent constructs a test measures or the outcomes it predicts best. For instance, tests of verbal ability and arithmetic reasoning predict academic achievement in both content realms about equally well, probably because both measure mostly the same latent ability, *g*. Likewise, different IQ tests tend to intercorrelate highly, indeed, often near the maximum possible given their reliabilities, despite often differing greatly in content and format (e.g., verbal vs. figural, group vs. individually administered, paper-and-pencil or not, short vs. long).

In ruling out superficial appearances as a guide to construct validity, the hierarchical model has simultaneously provided valuable new tools and tactics for determining what a test does and does not measure. Here is perhaps the most important. The *g* factors extracted from all broad batteries converge on the same "true" *g*, so this common *g* provides an external criterion against which to compare mental tests. Tests that correlate more strongly with *g* are more construct valid for assessing general intelligence regardless of appearance, label, or intent. All major IQ tests measure *g* well, yet not equally well. Some yield verbally flavored full-scale IQs, others perhaps spatially flavored IQs. This imprecision in measuring *g* matters not for most practical purposes (e.g., selection, placement, vocational counselling), but greatly for others (e.g., exploring the brain correlates of *g*).

The subtests of an IQ battery typically measure the general factor, *g*, to notably different degrees. Tests of memory and clerical speed are considerably less *g* loaded (more weakly correlated with *g*) than are tests of vocabulary, information, arithmetic reasoning, and matrix reasoning. Tests can likewise be compared in this manner against the different Stratum II factors, thusly indicating whether they tap the intended group factor (rather than *g* or something else), and how well they do so. In short, we can now place individual tests within the hierarchical structure, thusly indicating which construct(s) they actually measure and what we should and should not infer from scores on them. (How general vs. specific? If specific, which content domain?) This information can in turn be used to revise a battery or determine which subtests to include in a particular composite score. Any test can be evaluated using such tactics.

Finally, confirmatory factor analyses can be used to assess how well an entire battery replicates a prespecified model of intelligence, such as the CHC model. These analyses can lead to revising a test or how it is scored. For instance, the most recent edition of the Stanford–Binet provides four composite scores in place of its prior two in order to better fit the CHC model. This illustrates the iterative process by which the knowledge gleaned from tests about latent constructs can guide future revisions of those very same tests.

Just as the effort to understand intelligence has evolved from defining abilities to discovering them empirically, testing has evolved from validating tests primarily by strategy and intent during test development to marshalling postdevelopment networks of evidence to validate our interpretations of the behaviour they evoke.

### *Question 3: What Are Tests Good for, and Could They Be Made More Useful? Trend From Serving Mostly Institutional Needs to Serving Individuals Too*

We have already mentioned that intelligence tests were born of an operational need by mass institutions to assess, in a valid yet feasible way, the intellectual needs and talents of large populations. Group-administered cognitive ability tests are still widely used for selection and placement in civilian and military settings. They remain widely used because they provide useful information for organisational decision makers that no other source of information does, at least as accurately and economically.

Past successes have made us greedy for yet more detailed information about what individuals know and can do. For example, teachers want tests that allow them to track what their students still need to learn, say, in multiplying two-digit numbers (see lower half of Figure 1). That is, they want more formative feedback. Vocational counsellors likewise prefer ability profiles over general level of functioning in advising clients. And even school psychologists now shun global IQ scores, for diagnostic purposes, in favour of interpreting discrepancies across an individual's composite or subtest scores (their ability profile) or other specifics in test responses. The desire is understandable, but it remains to be seen whether discrepancy scores calculated across subtests or composites (upper right of Figure 1) reliably measure the latent traits (upper left) or specific achievements (lower half) of interest.

Some IQ test publishers have responded to user demand for more detailed assessments of individuals and more opportunity to exercise clinical judgement by deemphasizing global IQs in their technical manuals and reporting a greater variety of composite scores, usually comparable in breadth to Carroll's narrower Stratum II abilities or yet-narrower Stratum I abilities. Researchers are also working to produce new sorts of tests that provide more microlevel information, for instance, to gauge the specific cognitive processes that children use when learning and solving problems of different types. Large-scale cognitive diagnostic tests are an example (Leighton, 2009). As represented across the centre of Figure 1, these microlevel processes plumb further the depths of specificity, and perhaps approach the interface where aptitudes and achievements are so atomized that the distinction between them dissolves.

Whether any of these attempts to meet the new demand for increasing specificity actually provide valid information for users' intended purposes—whether they have treatment validity—has yet to be determined. Our point here is that this pressure from potential users does not reflect a demand for construct validity, but perhaps a growing discomfort about whether the constructs being measured, no matter how well, are the ones they want or know how to exploit. There seems to be a yearning for different, more manipulable, more democratically distributed individual differences. This may explain the attraction of more domain-specific tests and constructs, of tests of achievement rather than aptitude, and for assessing microlevel rather than global processes in cognitive processing. Clinicians are less interested in explaining and predicting performance than in opportunities to intervene constructively in the lives of particular individuals. The common presumption, also yet to be verified, is this requires more idiographic assessments at deeper levels of specificity and complexity.

Test publishing is a business and must respond to customers' needs and desires, and sometimes the dictates of third parties (such as insurers and regulatory agencies), but tests must also meet stringent technical and professional standards (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Society of Industrial and Organizational Psychology, 2003). The professional and political may pull in different directions, as just suggested. The major test publishers conduct sufficient research on their tests to satisfy professional test standards. We have described how construct validation transcends particular tests, and the test standards recognise that. Hence, anyone who claims their test is a valid measure of "intelligence" must give proper due to the broader body of scholarship on the construct and its measurement. Accordingly, the test manuals for several of the major intelligence batteries (e.g., WJ-III) now cite the CHC model as the consensus model of intelligence amongst scholars, and they describe how their test instantiates or is consistent with it. As just noted, the Stanford–Binet V doubled the number and changed the nature of the composite scores it reports to better reflect that model.

Efforts to satisfy conflicting masters are apparent, however. For instance, one well-known battery provides two sets of scores, one to reflect the CHC model and the other to reflect a competing model of intelligence. In addition, although intelligence test manuals now routinely cite the evidence for *g* as evidence for their battery's construct validity, they then deemphasize this crucial component of the battery and few report anything specific to it, such as the *g* loadings of the battery's subtests, composites, or overall IQ scores. One wonders whether an IQ battery's most construct-valid feature—its ability to measure general intelligence—has become its least attractive selling point. This point is addressed in the WAIS–IV and WISC–IV test manuals; although the FSIQ, reflecting *g*, is present in every analysis because of the cognitively complex nature of the test, it is the index scores that are argued to be the most useful for assisting differential diagnosis.

Intelligence tests can be useful without our knowing what constructs they measure, if any at all—for example, if we want to predict who will perform well if hired. The interest in gleaning more detailed profiles of individuals for diagnosis, counselling, and instruction is altogether different, however. It reflects not a disinterest in what tests measure, but a desire that they measure something fundamentally different—narrow abilities instead of broad ones or accomplishments rather their causes. A battery of tests can measure several things at once, but it cannot be made to measure opposing things at the same time. Intellectual assessment cannot be all things to all potential users.

To some extent, the desire for more detailed information is a desire specifically for information about developmental change. This is a different, more tractable matter in intellectual assessment. It requires an external criterion or common yardstick against which to compare individuals of different ages or the same person over time; IQ scores cannot provide that because they reset the average IQ to 100 at each age (are age-normed). Statistical methods are now available for creating a common yardstick for a given factor across different ages (e.g., for *g*, functional literacy, reading ability), although none can provide truly interval- or ratio-level measurement.

These trans-age scales are most useful, however, when they are behaviourally anchored—that is, when scale scores are tied to concrete examples of the sorts of tasks that individuals can routinely perform with proficiency (say, 70% probability of accurate response). Only a few standardised tests provide such interpretive options (e.g., the U.S. Department of Education's adult literacy surveys and its National Assessments of Educational Progress)—not even tests of achievement, which by their very nature are much more easily made *criterion-referenced.* To our knowledge, no intelligence test yet provides behavioural anchoring of scores, although it would add tremendously to their interpretability and diagnostic value. It would also help satisfy the shifting demands on tests, from serving primarily the needs of institutions (selection and placement) toward also serving the needs of practitioners and their clients (diagnosis, treatment, instruction, vocational counselling).

### *Question 4: Are Tests Fair, and Does Testing Enhance or Hurt the Larger Social Good? Trend Away From Eliminating Bias Toward Promoting Diversity*

From the earliest days of testing, the public and test developers alike have been concerned that tests assess people fairly and that scores not be artificially high or low owing to irrelevant or improper influences on test performance. Wechsler cautioned users of his earliest tests that their validity for assessing African Americans had not yet been established. There was rising concern in the 1960s, during the civil rights movement in the United States, that mental tests might be biased against Blacks and other minority groups and understate their abilities.

A large cadre of researchers began probing cognitive tests to determine whether they had equal predictive validity for all demographic groups, especially in school and work settings; whether their individual items functioned in the same manner for people of all racial and ethnic groups as well as both sexes; and whether either test administrators or test takers behaved in ways that artificially raised or lowered the scores of certain types of individuals. Measurement experts compared and debated different statistical definitions of bias and fairness; federal regulations and Supreme Court decisions in the United States ruled that differential passing rates by race would be considered prima facie evidence of unlawful discrimination, and began requiring proof of business necessity from employers if their tests were challenged for having adverse impact on minority groups. Simultaneously, there were lawsuits contesting the fairness of IQ testing for placement in special education owing to higher placement rates for African American children. One lawsuit resulted in a ban on the practise in California. College admissions testing has likewise been under political fire for decades.

These tests have been vetted for psychometric bias, as have their revisions, and all have passed (except where test takers are not native speakers of the language in which the test was administered, in which case a nonverbal test can be administered). Such vetting has become quite technically sophisticated. It ranges from examining all items individually (with item response theory techniques) to determining whether an entire battery yields the same factor structure or *construct invariance* for all demographic groups (using multiple-group confirmatory factor analysis). Potential items are also reviewed by stakeholders for appearance of unfairness. No test would meet the professional test guidelines or survive legal challenge today if it were psychometrically biased against (understated the abilities of) women or minority groups (see Weiss, Saklofske, Prifitera, & Holdnack, 2006). Of note is that, to the extent that tests understate abilities, the impact is on more able individuals in a population, and hence of groups with higher average scores (because imperfect reliability and validity always penalize more able individuals).

The good news about lack of cultural bias raises a new question on other grounds. It means that the persistent racial/ethnic gaps in average IQ and the profile differences by gender cannot be explained by mismeasurement. Regardless of their causes, they represent real disparities in average phenotypic or expressed and measured abilities, which in turn create group disparities in a wide variety of achievements, just like those ability differences amongst individuals within a group lead to unequal outcomes in the group.

The racial/ethnic gaps provoke the most contention partly because they have marked social consequences. When g-loaded tests are used to select, place, and promote individuals, race-blind use of test scores will produce racial imbalance in outcomes whenever members of different racial/ethnic groups differ in average level of *g*, as is the case in most operational settings. Degree of disparate impact is a highly predictable function of a test's *g* loading and the average IQ gap between the groups being assessed. Unfortunately, average racial gaps in IQ on valid, unbiased tests seem to be the rule, not the exception. For example, the average gap between White and both African American and Hispanic FSIQ scores on the WISC–IV FSIQ is 10 points (Weiss et al., 2006). Gaps may wax and wane somewhat, but are vexingly large and persisting (Jencks & Phillips, 1998; Jensen, 1998). This is one reason why some people argue that unbiased tests are not necessarily *fair* tests.

This conundrum—the inability to get race-blind and class-blind results from tests that are not biased by either race or class—has led to increasing pressure to modify tests or their manner of use to reduce disparate impact by race and class. Since the 1980s, the chief rationale for such modification has been to enhance diversity, the guiding presumption being that proportional representation by race and class within an organisation enhances all members' growth, well-being, and productivity.

Once again, external pressure has created somewhat inconsistent demands on test developers and test users. Supplementing a cognitive test with a noncognitive assessment typically increases predictive validity but does little to reduce disparate impact. Personnel psychologists have proved that there is no merely technical, scientific solution to the conundrum (Sackett, Schmitt, Ellingson, & Kabin, 2001; Schmitt, Rogers, Chan, Sheppard, & Jennings, 1997). Tests are not the fundamental problem; changing a test's content or format does not eliminate those differences or their causal impact on performance. Nor have other (non–*g*-related) domains of cognitive ability been overlooked; they do not exist. The only way to avoid disparate impact is to avoid measurements that reflect the groups' differences in *g* or to employ a multimethod approach to assessement and to use demographic factors to interpret test scores.

Whenever test takers from different races or classes differ substantially in *g*, the only way to substantially mitigate disparate impact in cognitive testing is to race-norm test scores (grade on a racial curve) or to use less reliable or less *g*-loaded mental tests, say, by reporting test scores in a only few broad categories (test-score banding), reducing the test's cognitive demands (its *g* loading), replacing it with subjective judgements of competence ("holistic" review), or switching to a personality inventory. One must greet with skepticism any claim to have found a special technique for eliminating most or all test score differences between groups that are known to differ on the latent trait (Gottfredson, 1996; Lohman, 2005; Lohman, Korb, & Lakin, 2008). As well, such strategies may well reduce a test's accuracy in order to increase diversity and therefore constitute sociopolitical decisions, not scientific or technical ones. In 1991, race-norming of employment tests was banned as a form of quota hiring. Some would argue, however, that such strategies improve *consequential validity.*

But that is our point. Neither the use of valid, unbiased cognitive tests nor a refusal to use them is a socially, economically, or politically neutral act. Nor can they ever be. Decisions taken in either direction affect how opportunities and resources will be distributed according to talent, achievement, and, incidentally, by race and class as well. Assessing the operational utility and social implications of gathering versus not gathering the information that tests provide requires making value judgements.

Decades back, test fairness was debated as a technical matter of test validity: Do tests measure latent constructs or predict outcomes equally accurately (with equal validity), regardless of social privilege, racial identity, or gender? Debates over fairness have shifted from disagreements over technical matters in measurement to disagreements over the social aims for testing—that is, from debating the validity of tests for specific purposes to debating their social consequences in a demographically diverse democratic society. Both are important debates, but only the former is a measurement issue. However, the latter is sometimes mistakenly framed as a technical matter. The shift toward framing social utility as if it were just another technical matter in adjudicating test validity is both embodied by and hidden in the notion of consequential validity, introduced in the 1980s. The wise psychologist using intelligence tests today will know that key factors such as affluence and education are highly correlated with FSIQ (Georgas et al., 2003), and that factors such as parent education, income, and expectations have reduced the WISC–IV FSIQ discrepancies to 6 points for Whites and African Americans and essentially 0 for White compared with Hispanic groups. So again, it is not the test but how we use it that is the issue here.

### *Question 5: Could Intelligence Tests Be Better Grounded in Knowledge of How the Mind and Brain Work? Trend Away From Debating Whether Psychometrics or Psychobiology Is the Best Approach to Understanding Intelligence Toward Joining the Two Approaches*

Enthusiasm is rising in some circles for the creation of "brain-based" cognitive tests, that is, ones that instantiate educational theories about how the typical brain learns. Their value is speculative because the theories are speculative. Psychometric tests are tools for capturing variance around the norm (individual differences), not for chronicling the developmental norm as it shifts over the life cycle, but variation does provide clues to causal processes. Cognitive variation can be correlated with individual differences in brain structure and function to understand which physiological attributes aid or impede proficient learning, reasoning, information retrieval, and other cognitive processes. Moreover, instruments are available to measure both psychometric and brain variation with precision. Such research is starting to burgeon (Haier, 2009; Jung & Haier, 2007) and, if cognizant of which aspects of the hierarchical structure a test is measuring, can provide valuable evidence about the biological instantiations of the various latent traits, better ways to measure them, and more meaningful interpretation and use of cognitive test results. These investigations are also starting to provide insight into sex differences in ability profiles (verbal vs. spatial).

This conjoining of psychometric and physiological approaches to assessing mental function reflects the two founding traditions in intelligence measurement: Binet's focus on accuracy in performing tasks that require complex information processing and Galton's focus on speed in reacting to exceedingly simple cognitive stimuli, called *elementary cognitive tasks*. The second, more reductionist, psychophysical approach was quickly pushed aside, but was revived later in the 20th century by scholars searching for the fundaments of psychometric *g* (Jensen, 2006). Their inspection time and choice reaction time tasks (where performance is measured in milliseconds) are neither psychometric nor physiological measures of mental acuity, but yet correlate with them at both the phenotypic and genetic levels.

Differences in *g* are so pervasively and consistently enmeshed in all aspects of brain function and cognitive performance that *g* may not be an ability but a property of the brain, such as overall processing efficiency. *g* is clearly not unitary at the physiological level, but likely a function of many metabolic, electrical, and other elemental processes. That would explain why *g* forms the psychometric core of all cognitive abilities, no matter how broad or narrow they are. Language and spatial visualisation are more localised in the brain, which allows their associated psychometric factors to vary somewhat independently. *g* is most certainly not unitary at the genetic level. Approximately a third of our genes are expressed in the brain, and current thinking amongst behaviour geneticists is that no single gene will account for more than a miniscule amount of the variance in normal intelligence.

Psychophysical measures of intellectual strength seem feasible in principle but unlikely in practise. But they do have one tremendous advantage that psychometric tests lack—ratio measurement. Speed (distance per time) is like height, weight, and many other physiological measures in that it has a zero point and is counted in equal units from there. No psychological scale can currently do that, either count from zero (total absence) or in equal units of quantity. Ratio measures of brain function might be exploited in some manner to provide more criterion-related or developmentally informative interpretations of psychometric tests.

Where once the psychometric and physiological traditions were viewed as rivals, the trend now is to see them as essential partners in understanding the cognitive competencies we aim to assess. This growing partnership might also provide a firm base for brain-based intellectual assessment that is both valid and useful. It might also help resolve the puzzle of the Flynn effect, which is the 3-point rise in IQ test scores in the United States per decade over most of the last century. Absent ratio scaling, it is difficult to know which component(s) of variance in IQ scores have increased over time— g itself, more specific abilities (e.g., scores have risen a lot on some subtests but not at all on others), or measurement artifacts. Although we do not need to understand the brain to measure intelligence, we cannot truly understand intelligence until we do.

## Summary

We have attempted to place intelligence and intelligence testing into a historical, theoretical, and evidence-based framework. Research and theoretical advances in cognitive psychology, neuropsychology, and developmental psychology continue to yield a wealth of new data about intelligence and cognitive processes and provide guideposts for both what we assess and how we assess it. Although controversy surrounding the assessment of intelligence is inevitable, the ever-increasing psychometric sophistication (e.g., inferential and continuous norming, Rasch scaling, differential item functioning, structural equation modelling) offers an increased capacity to measure intelligence with the sensitivity, precision, and economy of time required and expected by practitioners. A test is also continuously reviewed for its "capacity to perform" accurately. For example, when norms change (e.g., Flynn effect, shifting population demographics) or items simply "grow old," the test will require revision. And finally, the practise of psychology demands better clinical tests to aid in assessment, diagnosis, and intervention planning. With the given sophistication in knowledge and training, psychologists will use tests with professional and ethical integrity to help individuals and institutions accommodate cognitive diversity as, indeed, was the impetus for creating the very first intelligence test. As we have argued elsewhere (Gottfredson, 2008), "perhaps in no other applied setting is construct validity more important than for clinicians who are asked to diagnose individuals and intervene in their lives . . . arguably, a battery of cognitive tests is the most important single tool in sketching that portrait (p. 546)."

### Résumé

Aucun sujet n'est aussi central en psychologie que l'intelligence et sa mesure. Avec une histoire aussi ancienne que la psychologie en tant que telle, l'intelligence est le construit le plus étudié et peut-être le mieux compris en psychologie, même si plusieurs « questions sans réponses » subsistent. Le perfectionnement psychométrique des tests d'intelligence atteint des niveaux inégalés. Les auteurs font un survol de l'histoire, de la théorie et de la mesure de l'intelligence. Cinq questions portant sur des thèmes centraux à l'origine de confusion ou d'incompréhension associées à la mesure et l'évaluation de l'intelligence sont soulevées et discutées.

*Mots-clés* : intelligence, mesure de l'intelligence, tests d'intelligence

## References

Ackerman, P. L., & Beier, M. E. (2003a). Intelligence, personality, and interests in the career choice process. *Journal of Career Assessment, 11,* 205–218.

Ackerman, P. L., & Beier, M. E. (2003b). Trait complexes, cognitive investment and domain knowledge. In R. J. Sternberg & E. L. Grigorenko (Eds.), *Perspectives on the psychology of abilities, competencies, and expertise* (pp. 1–30). New York: Cambridge University Press.

Ackerman, P. L., & Kanfer, R. (2004). Cognitive, affective, and conative aspects of adult intellect within a typical and maximal performance framework. In D. Y. Dai & R. J. Sternberg (Eds.), *Motivation, emotion, and cognition: Integrated perspectives on intellectual functioning* (pp. 119–141). Mahwah, NJ: Erlbaum.

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.

Anastasi, A., & Urbina (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Boake, C. (2002). From the Binet-Simonto the Wechsler-Bellevue: Tracing the history of intelligence testing. *Journal of Clinical and Experimental Psychology, 24,* 383–405.

Bouchard, T. J., Jr. (1998). Genetic and environmental influences on adult intelligence and special mental abilities. *Human Biology, 70,* 257–279.

Bowden, S. C., Lange, R. T., Weiss, L. G., & Saklofske, D. H. (2008). Invariance of the measurement model underlying the WAIS–III in the United States and Canada. *Educational and Psychological Measurement, 68,* 1024–1040.

Bracken, B. A., & McCallum, R. S. (1998). *Universal Nonverbal Intelligence Test.* Itasca, IL: Riverside.

Canadian Psychological Association. (2000). *Canadian code of ethics for psychologists* (3rd ed.). Ottawa, ON: Author.

Carroll, J. B. (1993). *Human cognitive abilities.* New York: Cambridge University Press.

Collis, J. M., & Messick, S. (2001). *Intelligence and personality: Bridging the gap in theory and measurement.* Hillsdale NJ: Erlbaum.

Deary, I. J. (Ed.). (in press). Special issue on cognitive epidemiology. *Intelligence.*

Deary, I. J., Whalley, L. J., & Starr, J. M. (2009). A lifetime of intelligence. Washington, DC: American Psychological Association.

Eysenck, H. J. (1997). Personality and experimental psychology: The unification of psychology and the possibility of a paradigm. *Journal of Personality and Social Psychology, 73,* 1224–1237.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences.* New York: Basic Books.

Gardner, H., Kornbaber, M. L., & Wake, W. K (1996). *Intelligence: Multiple perspectives.* Ft. Worth, TX: Harcourt Brace College.

Georgas, J., Weiss, L. G., van de Vijver, F. J. R., & Saklofske, D. H. (2003). *Culture and children's intelligence: Cross-cultural analysis of the WISC–III*. San Diego, CA: Academic Press.

Gottfredson, L. S. (1996). Racially gerrymandering the content of police tests to satisfy the U.S. Justice Department: A case study. *Psychology, Public Policy, and Law, 2,* 418–446.

Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history and bibliography. *Intelligence, 24,* 13–23.

Gottfredson, L. S. (2004). Intelligence: Is it the epidemiologists' elusive "fundamental cause" of social class inequalities in health? *Journal of Personality and Social Psychology, 86,* 174–199.

Gottfreson, L. S. (2008). Of what value is intelligence? In Prifitera, A., Saklofske, D. H., & Weiss, L. G. (Eds.). *WISC-IV clinical assessment and intervention*. San Diego: Academic Press.

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.

Guilford, J. P. (1967). *The nature of human intelligence.* New York: McGraw-Hill.

Haier, R. J. (Ed.). (2009). Intelligence and the brain. *Intelligence, 37,* 121–230.

Herrnstein, R. J., & Murray, C. (1994). The bell curve: Intelligence and class structure in American life. New York: Free Press.

Hilgard, E. R. (1980). The trilogy of mind: Cognition, affection, and conation. *Journal of the History of Behavioral Sciences, 16,* 107–117.

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.

Jencks, C., & Phillips, M. (Eds.). (1998). *The Black–White test score gap.* Washington, DC: Brookings Institution.

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

Jensen, A. R. (2006). *Clocking the mind: Mental chronometry and individual differences.* New York: Elsevier.

Jung, R. E., & Haier, R. J. (2007). The parieto-frontal integration theory (P-FIT) of intelligence: Converging neuroimaging evidence. *Behavioral and Brain Sciences, 30,* 135–187.

Kaufman, A. S., Johnson, C. K., & Liu, Y. (2008). A CHC theory-based analysis of age differences on cognitive abilities and academic skills at age 22 and 90 years. *Journal of Psychoeducational Assessment, 26,* 350–381.

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

Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Assessment Battery for Children* (2nd ed.). Circle Pines, MN: American Guidance Service.

Kluckhohn, C., & Murray, H. A. (1948). *Personality in nature, culture and society.* New York: Knopf.

Lee, H. F., Gorsuch, R., Saklofske, D. H., & Patterson, C. (2008). Cognitive differences for ages 16 to 89 (Canadian WAIS–III): Curvilinear with Flynn and processing speed corrections. *Journal of Psychoeducational Assessment, 26,* 382–394.

Leighton, J. P. (2009). Mistaken impressions of large-scale diagnostic testing. In R. P. Phelps (Ed.), *Correcting fallacies about educational and psychological testing* (pp. 219–246). Washington, DC: American Psychological Association.

Lohmna, D. F., Korb, K. A., & Lakin, J. M. (2008). Identifying academically gifted English-language learners using nonverbal tests: A comparison of the Raven, NNAT, and CogAT. *Gifted Child Quarterly, 52,* 275–296.

Lohman, D. F. (2005). Review of Naglier and Ford 92003) Does the Naglieri Nonverbal Ability Test identify equal proportions of high scoring White, Black, and Hispanic Students? *Gifted Child Quarterly, 49,* 19–28.

McGrew, K. S., & Flanagan, D. P. (1998). *The intelligence test desk reference (ITDR): Gf-Gc cross-battery assessment.* Boston: Allyn & Bacon.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. K., Dies, R. R., et al. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56,* 128–165.

Naglieri, J. A. (2009). Intelligent intelligence testing: The influence of Alan S. Kaufman. In J. C. Kaufman (Ed.), *Intelligent testing: Integrating psychological theory and clinical practice* (pp. 73–98). New York: Cambridge University Press.

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

Priftera, A., Saklofske, D. H., & Weiss, L. G., (Eds.) (2008). WISC-IV clinical assessment and intervention. San Diego: Academic Press.

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., et al. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51,* 77–101.

Principles for Fair Student Assessment Practices for Education in Canada. (1993). Edmonton, AB: Joint Advisory Committee.

Roid, G. H. (2003). *Stanford–Binet Intelligence Scales* (5th ed.). Itasca, IL: Riverside.

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. *American Psychologist, 56,* 302–318.

Saklofske, D. H., & Zeidner, M. (1995). *International handbook of personality and intelligence.* New York: Plenum Press.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262–274.

Schmitt, N., Rogers, W., Chan, D., Sheppard, L., & Jennings, D. (1997). Adverse impact and predictive efficiency of various predictor combinations. *Journal of Applied Psychology, 82,* 719–730.

Schoenberg, M. R., Lange, R. T., & Saklofske, D. (2007). Estimating premorbid FSIQ for the Canadian WISC–IV: Demographic and a combined estimation procedures. *Journal of Clinical and Experimental Neuropsychology, 29,* 867–878.

Schoenberg, M. R., Lange, R. T., Saklofske, D. H., Suarez, M., & Brickell, T. A. (2008). Validation of the CPIE method of estimating premorbid FSIQ among children with brain injury. *Psychological Assessment, 20,* 377–384.

Society of Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.

Spearmon, C. (1904). “General Intelligence,” objectively defined and measured. *American Journal of Psychology, 15,*201–293.

Sternberg, R. (1997). *Successful intelligence.* New York: Plume.

Thorndike, E. L. (1924). Measurement of intelligence: 1. the present status. *Psychological Review, 31, 219.*

Thurstone, L. L. (1938). *Primary mental abilities.* Chicago: University of Chicago Press.

Tulsky, D. S., Saklofske, D. H., Chelune, G. J., Heaton, R. K., Ivnik, R. J., Bornstein, R., et al. (2003). *Clinical interpretation of the WAIS–III and WMS–III.* San Diego, CA: Academic Press.

Tzuriel, D. (2001). *Dynamic assessment of young children.* New York: Kluwer.

*Universal Declaration of Ethical Principles for Psychologists.* (2008). International Union of Psychological Science.

Vernon, P. E. (1950). The structure of human abilities. New York: Wiley.

Wechsler, D. (1939). *The measurement of adult intelligence.* Baltimore: Williams & Wilkins.

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

Wechsler, D. (1999). *The Wechsler Abbreviated Scale of Intelligence.* San Antonio, TX: The Psychological Corporation.

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

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

Wechsler, D., & Naglieri, J. A. (2006). *The Wechsler Nonverbal Scale of Ability.* San Antonio, TX: Harcourt Assessment.

Weiss, L. G., Saklofske, D. H., Prititera, A., & Holdnack, J. A. (2006). WISC-IV advanced clinical interpretation. San Diego: Academic Press.

Whipple, G. M. (1914). *Manual of mental and physical tests.* Baltimore: Warwick & York.

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

Received March 13, 2009
Revision received May 22, 2009
Accepted May 26, 2009 ■