PART I

# Fundamental Issues

# 1 Defining and Measuring Intelligence

## The Psychometrics and Neuroscience of *g*

Thomas R. Coyle

### Aims and Organization

The purpose of this chapter is to review key principles and findings of intelligence research, with special attention to psychometrics and neuroscience. Following Jensen (1998), the chapter focuses on intelligence defined as general intelligence (*g*). *g* represents variance common to mental tests and arises from ubiquitous positive correlations among tests (scaled so that higher scores indicate better performance). The positive correlations indicate that people who perform well on one test generally perform well on all others. The chapter reviews measures of *g* (e.g., IQ and reaction times), models of *g* (e.g., Spearman's model and the Cattell-Horn-Carroll model), and the invariance of *g* across test batteries.

The chapter relies heavily on articles published in the last few years, seminal research on intelligence (e.g., neural efficiency hypothesis), and meta-analyses of intelligence and *g*-loaded tests. Effect sizes are reported for individual studies and meta-analyses of the validity of *g* and its link to the brain.

The chapter is divided into five sections. The first section discusses historical definitions of intelligence, concluding with the decision to focus on *g*. The second section considers vehicles for measuring *g* (e.g., IQ tests), models for representing *g* (e.g., Cattell-Horn-Carroll), and the invariance of *g*. The next two sections discuss the predictive power of *g*-loaded tests, followed by a discussion of intelligence and the brain. The final section considers outstanding issues for future research. The issues include non-*g* factors, the development of intelligence, and recent research on genetic contributions to intelligence and the brain (e.g., Lee et al., 2018).

### Defining Intelligence

Intelligence can be defined as a general cognitive ability related to solving problems efficiently and effectively. Historically, several definitions of intelligence have been proposed. Alfred Binet, who co-developed the precursor

to modern intelligence tests (i.e., Stanford-Binet Intelligence Scales), defined it as "judgment, otherwise called good sense, practical sense, initiative, the faculty of adapting one's self to circumstances" (Binet & Simon, 1916/1973, pp. 42–43). David Wechsler, who developed the Wechsler Intelligence Scales, defined it as the "global capacity of the individual to act purposefully, to think rationally and to deal effectively with his environment" (Wechsler, 1944, p. 3). Howard Gardner, a proponent of the theory of multiple intelligences, defined it as "the ability to solve problems, or to create products, that are valued within one or more cultural settings" (Gardner, 1983/2003, p. x).

Perhaps the best known contemporary definition of intelligence was reported in the statement "Mainstream Science on Intelligence" (Gottfredson, 1997). The statement was signed by 52 experts on intelligence and first published in the *Wall Street Journal*. It defines intelligence as:

> [A] very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience. It is not merely book learning, a narrow academic skill, or test-taking smarts. Rather, it reflects a broader and deeper capability for comprehending our surroundings – catching on, making sense of things, or figuring out what to do. (Gottfredson, 1997, p. 13)

Two elements of the definition are noteworthy. The first is that intelligence represents a *general* ability, which influences performance on all mental tasks (e.g., verbal, math, spatial). The second is that intelligence involves the ability to learn *quickly*, meaning that intelligence is related to fast and efficient mental processing. Psychometric evidence strongly supports the view that intelligence, measured by cognitive tests, reflects a general ability that permeates all mental tasks, and that it is associated with efficient mental processing, notably on elementary cognitive tasks that measure reaction times (e.g., Jensen, 1998, 2006).

Arthur Jensen, a titan in intelligence research, advised against the use of the term "intelligence" because of its vague meaning and questionable scientific utility, noting: "Largely because of its popular and literary usage, the word 'intelligence' has come to mean too many different things to many people (including psychologists). It has also become so fraught with value judgments, emotions, and prejudices as to render it useless in scientific discussion" (Jensen, 1998, p. 48).

Rather than using the term "intelligence," Jensen (1998) proposed defining mental ability as *g*, which represents variance common to mental tasks. *g* reflects the empirical reality that people who do well in one mental task generally do well on all other mental tasks, a finding supported by positive correlations among cognitive tests. Following Jensen (1998), the current chapter focuses on *g* and *g*-loaded measures. It discusses methods for measuring *g*, models for representing *g*, the validity of *g*-loaded tests, and the nexus of relations between *g*-loaded measures, the brain, and diverse criteria.

## Measuring *g*

Jensen (1998, pp. 308–314) distinguished between constructs, vehicles, and measurements of *g*. The construct of *g* represents variance common to diverse mental tests. *g* is based on the *positive manifold*, which refers to positive correlations among tests given to representative samples. The positive correlations indicate that people who score high on one test generally score high on all others. *g* is a source of variance (i.e., individual differences) in test performance and therefore provides no information about an individual's level of *g*, which can be measured using vehicles of *g*.

Vehicles of *g* refer to methods used to elicit an individual's level of *g*. Common vehicles of *g* include IQ tests, academic aptitude tests (SAT, ACT, PSAT), and elementary cognitive tasks (ECTs) that measure reaction times. All mental tests are *g* loaded to some extent, a finding consistent with Spearman's (1927, pp. 197–198) *principle of the indifference of the indicator*, which states that all mental tests are loaded with *g*, irrespective of their content. A test's *g* loading represents its correlation with *g*. Tests with strong *g* loadings generally predict school and work criteria well, whereas tests with weak *g* loadings generally predict such criteria poorly (Jensen, 1998, pp. 270–294).

Measurements of *g* refer to the measurement scale of *g*-loaded tests (e.g., interval or ratio). IQ scores are based on an interval scale, which permits ranking individuals on a trait (from highest to lowest) and assumes equal intervals between units. IQ tests provide information about an individual's performance on *g*-loaded tests (compared to other members in his or her cohort), which can be converted to a percentile. Unfortunately, IQ scores lack an absolute zero point and therefore do not permit proportional comparisons between individuals such as "individual A is twice as smart (in IQ points) as individual B," which would require a ratio scale of measurement.

## *g*-Loaded Tests

This section reviews common tests of *g*. The tests include IQ tests, aptitude tests (SAT, ACT, PSAT), tests of fluid and crystallized intelligence, elementary cognitive tasks (ECTs), and tests of executive functions.

### IQ Tests

IQ tests include the Wechsler Intelligence Scales, which are among the most widely used IQ tests in the world. The Wechsler Scales are age-normed and define the average IQ at any age as 100 (with a standard deviation of 15). The scales yield four ability indexes: verbal comprehension, which measures verbal abilities (e.g., vocabulary knowledge); perceptual reasoning, which measures

non-verbal reasoning (e.g., building a model with blocks); processing speed, which measures psychomotor speed (e.g., completing a coding chart); and working memory, which measures the ability to manipulate information in immediate memory. Working memory is a strong correlate of *g* (e.g., Gignac & Watkins, 2015; see also, Colom, Rebollo, Palacios, Juan-Espinosa, & Kyllonen, 2004). Working memory can be measured using the Wechsler backward digit-span subtest, which measures the ability to repeat a series of digits in reverse order.

The Wechsler Scales yield a strong *g* factor (Canivez & Watkins, 2010), which measures variance common to mental tests. *g* largely explains the predictive power of tests, which lose predictive power after (statistically) removing *g* from tests (Jensen, 1998, pp. 270–294; see also, Coyle, 2018a; Ree, Earles, & Teachout, 1994). The Wechsler vocabulary subtest has one of the strongest correlations with *g* (compared to other subtests), suggesting that vocabulary knowledge is a good proxy of *g*. A Wechsler subtest with a relatively weak *g* loading is coding, which measures the ability to quickly complete a coding chart. Coding partly measures handwriting speed, which involves a motor component that correlates weakly with *g* (e.g., Coyle, 2013).

IQ scores are based on an interval (rather than ratio) scale, which estimates where an individual ranks relative to others in his or his age group. IQ scores can be converted to a percentile rank, which describes the percentage of scores that are equal to or lower than it. (For example, a person who scores at the 95th percentile performs better than 95% of people who take the test.) However, because IQ scores are not based on ratio scale (and therefore have no real zero point), they cannot describe where a person stands in proportion to another person. Therefore, IQ scores do not permit statements such as "a person with an IQ of 120 is twice as smart (in IQ points) as a person with an IQ of 60." For such statements to be meaningful, cognitive performance must be measured on a ratio scale. Reaction times are based on a ratio scale and do permit proportional statements (for similar arguments, see Haier, 2017, pp. 41–42; Jensen, 2006, pp. 56–58).

### Aptitude Tests

Aptitude tests are designed to measure specific abilities (verbal or math) and predict performance in a particular domain (school or work). Aptitude tests include the SAT and ACT, two college admissions tests taken in high school; the PSAT, a college readiness test taken in junior high school; and the Armed Services Vocational Aptitude Battery (ASVAB), a selection test used by the US military. All of these tests produce scores based on an interval scale and provide percentiles to compare an examinee to others in his or her cohort. The SAT, ACT, PSAT, and ASVAB also yield a strong *g* factor, which accounts for about half of the variance in the tests. All of the tests correlate strongly with IQ tests and *g* factors based on other tests, suggesting that they are in fact

"intelligence" tests, even though "intelligence" is not mentioned in their names. Finally, all of the tests derive their predictive power for work and school criteria largely (though not exclusively) from *g* (e.g., Coyle & Pillow, 2008; see also Coyle, Purcell, Snyder, & Kochunov, 2013).

### Tests of Fluid and Crystallized Intelligence

Cattell (1963; see also Brown, 2016; Horn & Cattell, 1966) distinguished between fluid intelligence, which measures general reasoning ability on novel problems, and crystallized intelligence, which measures culturally acquired knowledge. A widely used test of fluid intelligence is the Raven's Progressive Matrices. Each Raven's item depicts a 3 × 3 grid, with the lower right cell empty and the other cells filled with shapes that form a pattern. Participants must select the shape (from eight options) that completes the pattern. The Raven's correlates strongly with a *g* based on diverse tests (*g* loading [λ] ≈ .70), making it a good measure of *g*, and it also loads moderately on a visuospatial factor (λ ≈ .30, Gignac, 2015). Crystallized intelligence is often measured using vocabulary and general knowledge tests. Both types of tests measure culturally acquired knowledge and typically have among the highest correlations with a *g* based on diverse tests (λ ≈ .80, Gignac, 2015). Fluid and crystallized intelligence show different developmental trajectories over the lifespan (20–80 years). Fluid intelligence begins to decline in early adulthood and shows rapid declines in middle and late adulthood. In contrast, crystallized intelligence shows slight declines in later adulthood, with modest declines thereafter (e.g., Tucker-Drob, 2009, p. 1107).

### Elementary Cognitive Tasks (ECTs)

ECTs examine relations between *g* and mental speed using reaction times (RTs) to simple stimuli (e.g., lights or sounds) (for a review see Jensen, 2006, pp. 155–186). ECTs measure two types of RTs: Simple RT (SRT), which measures the speed of responding to a single stimulus (with no distractors), and choice RT (CRT), which measures the speed of responding to a target stimulus paired with one or more distractors. In general, RTs increase (become slower) with the number of distractors, which increases the complexity of the ECT. Moreover, RT-IQ relations, and RT relations with other *g*-loaded measures (e.g., working memory) increase as a function of task complexity. RT-IQ relations are weakest for SRT and stronger for CRT, with RT-IQ relations increasing with the complexity of the ECT (e.g., Jensen, 2006, pp. 164–166). Such a pattern is consistent with the idea that intelligence involves the ability to handle complexity (Gottfredson, 1997). A similar pattern is found when RT is correlated with participants' age in childhood (up to 20 years) or adulthood (20–80 years). Age correlates more strongly with CRT than with SRT, and

CRT relations with age generally increase with the number of distractors (e.g., Jensen, 2006, pp. 105–117).

ECTs can separate the effects of RT, which measures how quickly participants initiate a response to a reaction stimulus (light or sound), from movement time (MT), which measures how quickly participants execute a response after initiating it. RT and MT can be measured with the Jensen box (Jensen, 2006, pp. 27–31). The Jensen box involves a home button surrounded by a semicircle of one-to-eight response buttons, which occasionally light up. The participant begins with a finger on the home button, waits for a response button to light up, and then has to release the home button and press the response button. RT is the interval between the lighting of the response button and the release of the home button. MT is the interval between the release of the home button and the press of the response button. RT generally correlates more strongly with IQ and task complexity than does MT (Jensen, 2006, p. 234). Such results suggest that IQ reflects the ability to evaluate options and initiate a response (i.e., RT) more than the ability to execute a motoric response after deciding to initiate it (cf. Coyle, 2013).

### Executive Functions (EFs)

Executive functions are cognitive abilities used to plan, control, and coordinate behavior. EFs include three cognitive abilities: updating, which measures the ability to update information in working memory; shifting, which measures the ability to shift attention to different stimuli or goals; and inhibition, which measures the ability to suppress distractions (Miyake et al., 2000). Of the three EFs, updating and its analog of working memory correlate most strongly with *g* (e.g., Friedman et al., 2006; see also, Benedek, Jauk, Sommer, Arendasy, & Neubauer, 2014). The relation between working memory and *g* approaches unity in latent variable analysis (e.g., Colom et al., 2004; see also, Gignac & Watkins, 2015), with a mean meta-analytic correlation of .48 among manifest variables (Ackerman, Beier, & Boyle, 2005). The three major EFs (updating, shifting, inhibition) are related to each other, suggesting a general EF factor. Controlling for correlations among the three EFs indicates that updating (an analog of working memory) correlates most strongly with *g*, whereas shifting and inhibition correlate weakly with *g* (e.g., Friedman et al., 2006).

## Models of Intelligence and *g*

Two prominent models of *g* are a Spearman model with no group factors, and a hierarchical model with group factors (Jensen, 1998, pp. 73–81). Group factors estimate specific abilities (e.g., verbal, math, spatial), whereas *g* estimates variance common to all abilities. Group factors (and the tests used to estimate them) almost always correlate positively, reflecting shared variance among the factors. The Spearman model estimates *g* using manifest variables

(e.g., test scores), with no intervening group factors. In contrast, the hierarchical model estimates *g* based on a pyramidal structure, with *g* at the apex, group factors (broad and narrow) in the middle, and manifest variables (individual tests) at the base.

There are many hierarchical models of *g* with group factors. One of the most notable is the Cattell-Horn-Carroll (CHC) model (McGrew, 2009). The CHC model describes *g* as a third-order factor, followed by broad second-order group factors, each loading on *g*, and narrow first-order group factors, each loading on a broad factor. The broad factors (sample narrow factors in parentheses) include fluid intelligence (induction), crystallized intelligence (general knowledge), quantitative knowledge (math knowledge), processing speed (perceptual speed), and short- and long-term memory (working memory capacity). In practice, intelligence research often targets *g* and a small number of group factors relevant to a study's aims. It should be emphasized that all group factors (broad and narrow) are related to *g*. Therefore, the unique contribution of a group factor (e.g., math ability) to a criterion (e.g., school grades) can be examined only after statistically removing *g* from the factor, a point revisited in the section on non-*g* factors.

## Invariance of *g*

Using hierarchical models of *g*, Johnson and colleagues (Johnson, Bouchard, Krueger, McGue, & Gottesman, 2004; Johnson, te Nijenhuis, & Bouchard, 2008) estimated correlations among *g* factors based on different batteries of cognitive tests. An initial study (Johnson et al., 2004; $N = 436$ adults) estimated *g* and diverse group factors using three test batteries: Comprehensive Ability Battery (14 tests estimating five group factors), Hawaii Battery (17 tests estimating five group factors), and Wechsler Adult Intelligence Scale (11 tests estimating three group factors). *g* factors for each battery were estimated as second order factors in latent variable analyses. Although the three batteries differed on key dimensions (e.g., number of tests, content of tests, number of group factors), the *g* factors of the batteries correlated nearly perfectly ($r \approx 1.00$). The near perfect correlations suggest that *g* is independent of specific tests and that *g* factors based on diverse test batteries are virtually interchangeable.

Johnson et al.'s (2004) results were replicated in a subsequent study (Johnson et al., 2008), cleverly titled "Still just 1 *g*: Consistent results from five test batteries." The study involved Dutch seamen ($N = 500$) who received five test batteries. The batteries estimated *g* and different group factors (perceptual, spatial, mechanical, dexterity), with few verbally loaded factors. Consistent with Johnson et al.'s (2004) results, the *g* factors of the different batteries correlated .95 or higher, with one exception. The exception was a test battery composed entirely of matrix type reasoning tests (Cattell Culture Fair Test), which yielded a *g* that correlated .77 or higher with the *g* factors of the other

tests. In Johnson et al.'s (2008) words, the results "provide evidence both for the existence of a general intelligence factor [i.e., *g*] and for the consistency and accuracy of its measurement" (p. 91).

Johnson et al.'s (2004, 2008) results are consistent with Spearman's (1927) principle of the indifference of the indicator. This principle is based on the idea that all cognitive tests are indicators of *g* and load on *g* (to some extent). The *g* loading of a test represents its correlation with *g*, which reflects how well it estimates *g*. The degree to which a test battery estimates *g* depends on the number and diversity of tests in the battery (e.g., Major, Johnson, & Bouchard, 2011). Larger and more diverse batteries like the ones used by Johnson et al. (2004, 2008) generally yield better estimates of *g* because such batteries are more likely to identify variance common to tests (i.e., *g*) and have test-specific variances cancel out.

Johnson et al.'s (2004, 2008) research estimated *g* using samples from WEIRD countries (e.g., United States and Europe). WEIRD stands for Western, Educated, Industrialized, Rich, and Democratic. WEIRD countries have high levels of wealth and education (Henrich, Heine, & Norenzayan, 2010), which contribute to cognitive development, specific abilities (e.g., verbal, math, spatial), and *g*. Non-WEIRD countries have fewer resources, which may retard cognitive development and yield a poorly defined *g* factor (which explains limited variance among tests). Warne and Burningham (2019) examined *g* factors in non-WEIRD countries (e.g., Bangladesh, Papua New Guinea, Sudan). Cognitive test data were obtained for 97 samples from 31 non-WEIRD countries totaling 52,340 individuals. Exploratory factor analyses of the tests estimated *g*, defined as the first unrotated factor when only one factor was extracted, or a second-order factor when multiple factors were extracted. A single *g* factor was observed in 71 samples (73%), and a second-order *g* factor was observed in the remaining 23 of 26 samples (83%). The average variance explained by the first unrotated factor was 46%, which is consistent with results from WEIRD countries. In sum, a clearly identified *g* factor was observed in 94 of 97 non-WEIRD countries, suggesting that *g* is a universal human trait, found in both WEIRD and non-WEIRD countries.

## Predictive Power of *g* and *g*-Loaded Tests

Intelligence tests are useful because they predict diverse criteria in everyday life. The current section reviews research on the predictive power of intelligence at school and work. The review focuses on recent and seminal studies of *g*-loaded tests. *g*-loaded tests include IQ tests (Wechsler Scales), college aptitude tests (SAT, ACT, PSAT), military selection tests (Armed Services Aptitude Battery), and other cognitive tests (e.g., ECTs). In general, any test that involves a mental challenge will be *g*-loaded (to some extent), with the degree of relatedness between a test and *g* increasing with task complexity.

## Intelligence and School

Intelligence tests were developed to predict school performance and so it is no surprise that they predict school grades. Roth et al. (2015) examined the meta-analytic correlation between intelligence tests (verbal and nonverbal) and school grades with 240 samples and 105,185 students. The population correlation was .54 after correcting for artifacts (measurement error and range restriction). Moderating analysis indicated that the test-grade correlations increased from elementary to middle to high school (.45, .54, .58), and were stronger for math/science (.49) than for languages (.44), social sciences (.43), art/music (.31), and sports (.09). Roth et al. (2015) argued that the increases in effect sizes across grade levels could be attributed to increases in the complexity of course material, which would decrease the ability to compensate with practice and increase the contribution of intelligence.

Are intelligence-grade correlations attributable to students' socioeconomic status (SES), which reflects parental wealth, education, and occupational status? The question is important because intelligence tests and college admissions tests (SAT) have been assumed to derive their predictive power from SES. To address this question, Sackett, Kuncel, Arneson, Cooper, and Waters (2009) meta-analyzed SAT-GPA correlations using college GPAs from 41 institutions and correcting for range restriction. (The SAT correlates strongly with a *g* based on diverse tests [$r = .86$, corrected for nonlinearity, Frey & Detterman, 2004].) The meta-analytic SAT-GPA correlation was .47, which dropped negligibly to .44 after controlling for SES (Sackett et al., 2009, p. 7). Contrary to the assumption that the SAT derives its predictive power from SES, the results suggest that SES has a negligible impact on SAT-GPA correlations.

The predictive power of admissions tests is not limited to undergraduate criteria but also applies to graduate and professional school criteria. Kuncel and Hezlett (2007) meta-analyzed correlations involving graduate admissions tests, correcting for range restriction and measurement error. The tests included the Graduate Record Examination (GRE), Law School Admission Test (LSAT), Pharmacy College Admission Test (PCAT), Miller Analogies Test (MAT), Graduate Management Admission Test (GMAT), and Medical College Admission Test (MCAT). The tests robustly predicted first-year graduate GPA ($r > .40$, all tests), overall graduate GPA ($r > .40$, all tests), and qualifying exams ($r > .39$, GRE and MAT). Moreover, the tests also predicted criteria other than grades, including publication citations ($r \approx .23$, GRE), faculty evaluations ($r > .36$, GRE and MAT), and licensing exams ($r > .45$, MCAT and PCAT). All correlations were positive, indicating better performance was associated with higher achievement.

Are test–grade correlations attributable to *g*? The question is important because *g* is considered the "active ingredient" of tests, with the predictive power of a test increasing with its *g* loading. To address this question, Jensen