Conceptualizing and Measuring Intelligence

(A version of which to be published in the *SAGE Handbook of Personality and Individual Differences*

Accepted December 6th, 2016)

Gilles E. Gignac

University of Western Australia

Gignac, G. E. (in press). Conceptualizing and measuring intelligence. In V. Zeigler-Hill & T. Shackelford (Eds.), *The SAGE handbook of personality and individual differences*. Thousand Oakes, CA: Sage.

Author Note

Correspondence should be addressed to Gilles E. Gignac, School of Psychology, University of Western Australia, 35 Stirling Highway, Crawley, Western Australia, 6009, Australia. E-mail: gilles.gignac@uwa.edu.au

Abstract

Intelligence may be defined as an entity's maximal capacity to achieve a novel goal successfully using perceptual-cognitive abilities. More operationally, psychometric intelligence may be defined as an entity's maximal capacity to complete a novel, standardized task with veridical scoring using perceptual-cognitive abilities. With these definitions in mind, the following four key dimensions of intellectual intelligence are reviewed in relation to their nature and measurement: memory span, crystallized intelligence, reasoning, and processing speed. The nature and measurement of general intelligence is then also discussed. Finally, the prospect of the brief measurement of general intelligence is evaluated critically. It is concluded that knowledge relevant to our understanding of cognitive abilities has accumulated much more impressively over the last couple of decades, in comparison to advancements in the measurement of cognitive abilities. Research devoted to the advancement of cognitive ability measurement is encouraged.

Keywords: fluid intelligence; crystallized intelligence, processing speed, general intelligence

Conceptualizing and Measuring Intelligence

The measurement of any psychological construct requires a precise and unambiguous definition of the construct (Messick, 1981; Slaney & Racine, 2013). Because there are many different definitions of intelligence, even precise and relatively unambiguous ones, there are also many different approaches to the measurement of intelligence. In this chapter, I will focus only upon intellectual intelligence, as distinct from emotional and social intelligence, for example.

Many putative definitions of intelligence actually consist of a delineation of dimensions of intelligence, rather than a genuine definition of intelligence. For example: the ability to recall items from memory, the ability to reason, and the capacity to process simple instructions quickly is not a definition of intelligence. Although these dimensions of intelligence may be important in the discussion of the nature of intelligence, they do not constitute a definition of intelligence. A genuine definition of a construct, in this case intelligence, needs to represent what all the dimensions of interest are hypothesized to have in common (Nunnally, 1978).

To help overcome many of the misunderstandings and disagreements in the literature, Eysenck (1986, 1988) distinguished usefully between biological intelligence, psychometric intelligence, and practical intelligence. However, other than to make reference to well-known test batteries, Eysenck does not appear to have defined psychometric intelligence with precise terms. Lohman (1989) defined psychometric intelligence as "…that intelligence which is measured by psychometric tests" (p. 351). This definition is unsatisfactory, because the well-known Rosenberg Self-Esteem Questionnaire is a psychometric test (Rosenberg, 1965), but it should certainly not be considered an intelligence test. Additionally, the Intelligence Questionnaire is a nine-item self-report inventory designed to measure perceptions of ability across a number of domains widely considered indicative of intellectual functioning (Gignac, Stough, & Loukomitis, 2004), but should not be considered an intelligence test, either.

In my view, the three key characteristics that bind the commonly discussed dimensions of psychometric intelligence include: (1) maximal performance (Cronbach, 1960), (2) relatively novel problems or goals (Davidson & Downing, 2000; Raaheim & Brun, 1985), and (3) perceptual-cognitive processes (Thomson, 1919). Consequently, I define intelligence as an entity's maximal capacity to achieve a novel goal successfully using perceptual-cognitive abilities. The above definition does not imply that all, or even any, goals can be achieved solely through the application of perceptual-cognitive abilities. Undoubtedly, the achievement of virtually any goal would be affected by additional attributes (e.g., motivation, opportunity, personality). However, those other attributes would likely not be wholly perceptual-cognitive in nature, consequently, they should not be categorized as intellectual abilities (e.g., emotional intelligence would involve substantial emotional characteristics).

The definition of intelligence I provided above is relatively abstract, purposely so. The definition avoids the rather narrow conception of intelligence as necessarily relevant to tests or testing. Thus, I would argue that intelligence is not what intelligence tests measure, as originally stated by Boring (1923) and commonly repeated (e.g., Neisser, 1979; Orton, 1986; van der Maas, Kan, & Borsboom, 2014). Many novel goals were achieved, many novel problems solved, using perceptual-cognitive abilities prior to the publication of the first purpose developed intelligence test in 1905, the Binet-Simon Scale (Binet & Simon, 1916).

Despite the theoretical appeal associated with an abstract definition of a construct (Priem & Butler, 2001), abstract definitions are associated with limited scientific usefulness, as they lack a sufficient amount of concreteness to generate clear possibilities for measurement, not to mention hypotheses that can be tested empirically. For example, researchers and clinicians cannot be expected to follow people around waiting to see when they will apply their maximal perceptual-cognitive abilities to reach a self-generated goal that is novel. Consequently, a compromise of sorts has to be made, in order to obtain a definition of the construct that is more useful scientifically. I define psychometric intelligence *operationally* as an entity's maximal capacity to complete a novel, standardized task with veridical scoring using perceptual-cognitive abilities.

Thus, the psychometric inventories reported above (i.e., the Rosenberg Self-Esteem Inventory and the Intelligence Questionnaire) are not measures of intelligence, because they do not involve the successful completion of a task for which there is veridical scoring. Veridical scoring refers to the evaluation of a response with an objective and verifiably determined scoring key. For example, if a person were asked to recall the following sequence of digits, 4, 7, 2, 8, 3, 5, the objective and verifiably correct response would involve the recollection and expression of the said six digits. There is not debate and there is no need to solicit the opinion of experts (i.e., consensual scoring).

Memory Span: Short-Term

Memory span may be defined as the maximum capacity of sequential information an individual can recall accurately (Gathercole, 1999; Dehn, 2008). In more operational terms, memory span is the length of a series of objects for which the probability of immediate reproduction following a single presentation is fifty percent (Watkins, 1977). That is, for which there is an even chance of perfect recall. It is useful to start with a discussion about the measurement of psychometric intelligence with the topic of the measurement of memory span for two reasons. First, the typical measurement of memory span can satisfy the criteria for genuine scientific quantification (Michell, 1997). Secondly, the importance of memory span as a valid and socially valuable indicator of intellectual functioning appears to have been widely accepted, in comparison to other commonly measured cognitive ability dimensions (e.g., reasoning, processing speed, verbal comprehension), which have been the source of more substantial criticism.

There are many memory span dimensions that have been proposed over the years. In fact, Tulving's (2007) review of the literature suggests that there are over 250. In the context of cognitive ability testing, researchers typically administer tests to evaluate short-term memory, rather than long-term memory. As the name implies, short-term refers to maintaining objects in memory for a brief period of time. How brief? In practical terms, short-term memory tests typically present the objects (say, digits) for approximately 1 second each. The total testing time for any particular sequence of objects is less than 30 seconds. The participant must recall the objects immediately after the presentation of the last object associated with the sequence.

Most commonly, the objects selected for administration are single digits within a sequence (e.g., Wechsler, 2008; Williams, 1991). The psychometric properties associated with digit span, letter span, and single-syllable word span are similar ($\alpha \approx .80$; correlation with Raven's = .33 to .40; Kane et al., 2004). However, many other types of stimuli can be used as objects. For example, multi-syllable words (e.g., Maehler & Schuchardt, 2009), musical notes (e.g., Williamson, Baddeley, & Hitch, 2010), and visual shapes (e.g., Gonzalez, Thomas, & Vanyukov, 2005). It is worth noting that there is some evidence to suggest that visual span and spatial span may not be isomorphic constructs (Vicari, Bellucci, & Carlesimo, 2003). Spatial span is more relevant to memory for one or more objects' location in space (Corsi Blocks; Corsi, 1972; Dang, Braeken, Ferrer, & Liu, 2012), whereas visual span is memory for one or more objects' nature (e.g., Rudkin, Pearson, & Logie, 2007). A comprehensive latent variable modeling analysis relevant to the possible distinction between spatial and visual span appears to be lacking in the literature.

With respect to digits, the typical healthy adult can recall 6.56 digits (Gignac, 2015). There is relatively little variability in the amount of objects recalled across healthy adults (*SD* = 1.22). Specifically, approximately 80% of the population can recall between 5 and 8 objects in short-term memory. The difference amounts to only three objects, however, the implications are substantial, as the correlation between digit span and years of education completed has been reported to be moderate in magnitude, for example (e.g., *r* = .44; Paul et al., 2005).

Although single digits are most commonly used as stimuli in the psychological assessment of memory span, there is evidence to suggest that the use of letters, words, and visual objects can yield respectable levels of reliability and validity, as well (e.g., Kane et al., 2004). The degree to which the various modalities tap both common and unique memory span processes is an active area of research (e.g., Fougnie, Zughni, Godwin, & Marois, 2015; Giofrè, Mammarella, & Cornoldi, 2013). Although the results are not entirely consistent, the research to-date supports the notion that there is a fairly substantial general memory span process, as well as smaller, specific processes unique to each modality (Li, Christ, & Cowan, 2014).

Short-term memory span ability permeates virtually all other cognitive abilities. For example, it has been demonstrated that the solution of arithmetic problems involves the use of short-term memory (Noël, Désert, Aubrun, & Seron, 2001). Perhaps more surprisingly, a connection between short-term memory capacity and individual differences in vocabulary acquisition has also been argued based on theoretical and empirical evidence (Brown & Hulme, 1996; Gathercole, 2000). It is likely due to short-term memory's ubiquitous nature that serious deficits in memory span (e.g., Alzheimer’s disease) can have such profound impairments in a person’s day-to-day living.

From a more technical perspective, one of the limitations associated with the typical administration of a memory span test is the use of a discontinue rule. A discontinue rule demarcates the termination of testing after a specified number of erroneous responses. Digit Span within the Wechsler scales, for example, recommends the employment of a discontinue rule after two successive errors, irrespective of whether the trials are across items (e.g., 5 digit sequence and then 6 digit sequence). The consequence of employing such a discontinue rule is to bias internal consistency upwardly. Additionally, there may be a reduction in validity, because a person may be performing near their 50% chance of successful performance, but simply become distracted on the second trial. Woods et al. (2011) used a computer adaptive approach to digit span administration and observed enhanced validity in test scores, in comparison to the conventional approach to digit span administration. The advantages of less testing time and more accurately estimated point-estimates make computer adaptive testing in memory span research attractive. More research and developments in this area would be of great benefit.

It is also worth pointing out that, according to the WAIS-IV normative sample (Wechsler, 2008), approximately 10% of individuals in the normal population can achieve the highest score possible (i.e., recall of 9 digits). Thus, there is a ceiling effect, one which is exacerbated in high ability samples (e.g., university students). The inclusion of an additional digit sequence of 10 digits is a worthwhile addition, in order to increase the amount of variability in the data.

An old and commonly articulated argument against the contention that memory span is a good quality indicator of intellectual ability is that ordinary people can increase their memory span greatly through practice (Martin & Fernberger, 1929). As discussed next, such an argument is misguided, as it fails to recognize the important distinction between intelligence and expertise.

Distinguishing Intelligence from Expertise

One of the key attributes associated with the definition of intelligence provided above is the concept of novelty. In the context of intelligence and its measurement, novelty refers to a question, task, or problem that has not been experienced previously. Thus, in order for test scores to have respectable validity as indicators of intellectual functioning, the participants must not have specific knowledge about the nature of the test items prior to administration of the test (Jensen, 1998) and they certainly must not have practiced solving problems similar to those contained within the cognitive ability test(s).

There are many cases in the literature where individuals have practiced cognitive ability type tasks deliberately and conscientiously. Over time, these individuals have gained a capacity to achieve remarkably high scores on those specific tests. Perhaps most notably, two “ordinary” individuals were reported by Chase and Ericsson (1982) to have achieved single digit memory spans of 68 and 82 digits, respectively, after approximately 250 hours of practice over the course of two years. Such scores are 10 times higher than typically observed in healthy adults. Some may contend that such an observation vitiates the argument that there are natural (biological) individual differences which facilitate memory span capacity and intellectual functioning more broadly. However, it is important to re-state the characteristic of novelty in the context of valid intelligence testing. As the two individuals in the Chase and Ericsson (1982) study practiced the task of recalling single digits over many hours, novelty was no longer present. Thus, although the two individuals may have achieved remarkably high scores on a digit span test, it would not imply that the two individuals have a remarkable memory span capacity. Instead, it would be more accurate to contend that the two participants developed a specific expertise at completing a task, rather than enhanced their memory span cognitive ability.

In the context of mental abilities, expertise may be defined as a consistent superior cognitive skill acquired by repeatedly performing a task (Anderson, 2005). Based on extensive experimental and computational research, it is likely that memory span experts develop a mnemonic strategy known as chunking to achieve their impressive memory span scores (Chase & Simon, 1973; De Groot & Gobet, 1996; Richman, Staszewski, & Simon, 1995). Chunking involves the conscious organization of stimuli into groups to facilitate the storage and recall of information from short-term memory. Thus, with a mnemonic strategy such as chunking, the vast majority of healthy individuals would be expected to demonstrate the skill of achieving high memory span scores, but not an increase in their memory span capacity.

To help support such an argument, consider that chess experts have been demonstrated to recall board positions with a high level of accuracy (i.e., all 24 pieces), whereas novices can only recall the position of 4 pieces (Chase & Simon, 1973). Importantly, however, the chess experts' apparent exceptional memory span capacity was non-existent when the chess pieces were placed onto the board in random formations. Thus, the chess experts' impressive skill was highly circumscribed, rather than in any way generalized. Expertise in memory span is not the only one that has been found to be highly specific. Jensen (1990) tested Shakuntala Devi, a prodigious mental calculator, across a number of cognitive ability tests (e.g., Raven's, reaction time) and found him to score in the average range. Perhaps most surprisingly, Shankuntala Devi was found to achieve a digit span forward and backward scores of 9 and 4 digits, respectively (i.e., normal range). Thus, extensive practice at arithmetic did not translate into a superior memory span capacity for digits.

Some have argued that expertise is essentially a monotonic function of time spent practicing, known as the deliberate practice perspective (Ericsson, Krampe, & Tesch-Romer, 1993; Lewandowsky & Thomas, 2009). By contrast, others have contended that there are natural individual differences which affect the degree to which individuals can develop expertise, known as the talent (intelligence) perspective (Plomin & Petrill, 1997). It should be emphasized that the talent perspective does not deny that practice is required to achieve expert performance. Instead, the talent perspective proposes that, for most behaviors, substantial individual differences will be observed in the speed at which expertise can be developed, in addition to the maximal expert capacity that can be achieved. Furthermore, it may also be the case that deliberate practice is only necessary but not sufficient condition for the achievement of expertise, from the latent perspective. Thus, not all individuals would be expected to achieve expertise at all, or even necessarily any, particular skills, no matter how much they practice.

Until relatively recently, there was a paucity of impressive heritability research on the topic of expertise, thus, a convincing case for the talent (intelligence) perspective was lacking. However, Plomin, Shakeshaft, McMillan, and Trzaskowski (2014) investigated reading skill as an expertise amongst a sample of 10,000 12-year-old twins. Experts were defined as those students who performed at the top 95th percentile and above. Plomin et al. (2014) found that a MZ twins had a 69% chance of reading expertise concordance. By comparison, DZ twins had only a 38% chance of reading expertise concordance. Based on a number of different approaches to analyzing the data, Plomin et al. (2014) found that approximately 60% of the variance in reading expertise was heritable. Perhaps the key challenge associated with the deliberate practice view is that the same amount of practice does not necessarily lead to the same amount of expert performance.

Why do those individual differences exist? It has been found that only about half of the variance in chess expertise can be accounted for by deliberate practice (Charness, Krampe, & Mayr, 1996). Furthermore, there are substantial individual differences in the amount of practice required to achieve expert status in chess. Specifically, in a sample of 104 chess players, Gobet and Campitelli (2007) found that the relatively slowest expert chess achiever required eight times more practice than the relatively fastest expert chess achiever. In another study, Mosing, Madison, Pedersen, Kuja-Halkola, and Ullén (2014) administered the Swedish Musical Discrimination Test to a sample of 2,569 twin pairs. The twins were also asked how much time they had spent practicing in their lifetime across several categorical age bands (from childhood to adulthood). Importantly, Mosing et al. (2014) found that MZ twins who differed in the amount of practice they engaged in their lifetime did not differ statistically significantly in their music ability, as measured by the SMDT. These results suggest that there is a limit that practice can have on skill/expertise acquisition, which supports the talent perspective to expertise. Thus, individual differences in the rate at which people can learn, and/or gain expertise, is reflective of intellectual capacity.

Memory Span: Working Memory Capacity

Historically, short-term memory has played a relatively minor role in the assessment of intelligence, as testified by the fact that the Wechsler scales included only one subtest of memory span (Digit Span) for many years (Gignac & Weiss, 2015). In more recent times, individual differences in working memory has enjoyed a substantial amount of attention in the research and assessment community. In fact, it has even been suggested that working memory capacity may be the foundation of intellectual functioning (Colom, Rebollo, Palacios, Juan-Espinosa, & Kyllonen, 2004; Kyllonen & Christal, 1990).

Theoretically, the key distinguishing feature of working memory is the simultaneous maintenance and manipulation of information in memory, rather than simply the passive maintenance of information in memory (Baddeley & Hitch, 1974). From a differential psychology perspective, working memory is typically considered, at least partially, distinct from short-term memory (Cantor, Engle & Hamilton, 1991). Perhaps one of the simplest indicators of working memory capacity is digit span backward. Because digit span backward requires participants to recall the digits in the reverse order with which the digits were presented, it has been argued that some additional mental manipulation is required to execute the task successfully, in comparison to digit span forward. The extra manipulation has been suggested to explain the observation that digit span backward is a better indicator of general intellectual functioning (Gignac & Weiss, 2015; Oberauer, Süß, Schulze, Wilhelm, & Wittmann, 2000).

Interestingly, the correlation between digit span forward and digit span backward, two ostensibly similar tests, has been reported to be rather moderate at .55, based on the WAIS-IV normative sample (N = 2,200; Wechsler, 2008). The rather low correlation between the two similar tests raises the question of whether they should be combined together into a single composite (Reynolds, 1997). The Wechsler scales now provide percentile information for both forward and backward digit span, separately. Arguably, such information should be used by clinicians. To reflect the increased awareness of the value of working memory as an indicator of intellectual functioning, the Wechsler scales have added two additional subscales of working memory: letter number sequencing and digit span sequencing. Confirmatory factor analytic research reported by Gignac and Weiss (2015) suggests quite clearly that digit span backward and letter-number sequencing would make a much more justifiable composite indicator of working memory capacity, in comparison to the current recommendation of Digit Span (forward and backward combined) and Arithmetic.

Although digit span backward is a commonly regarded indicator of working memory capacity, it is arguably limited. Conway, Getz, Macnamara, and Engle de Abreu (2010) distinguished qualitatively between two types of working memory tasks: (1) coordination/transformation tasks and (2) complex span tasks. Coordination/transformation tasks include the previously described digit span backward, letter-number sequencing, and digit span sequencing type tests. Such tasks only require a transformation of order to complete successfully. By contrast, complex span tasks are more explicitly dual-task in nature, as they more clearly require processing and storage to complete successfully. Perhaps the most well-regarded complex span measurement paradigm is the *n*-back approach. In a typical *n*-back task, a series of stimuli are presented on a screen (e.g., individual letters). As the sequence of stimuli are presented on the screen, the participant is required to identify whether a visual stimulus does or does not match the one from n-steps earlier in the sequence, where n-steps is specified to the participant prior to the sequence run. For each stimulus, the participant must provide a response by depressing a key on a key board (non-target button or target button). In a 0-back task, the participant need only identify whether a presented letter matches a pre-determined target (e.g., letter ‘k’). In a 1-back task, the participant is required to identify when a letter has been presented on the screen two times in succession. The 0-back and 1-back tasks are considered principally measures of attention, rather than working memory, however, they are often administered in neuropsychological and brain imaging studies to facilitate the measurement of highly valid (i.e., comparable) control conditions (e.g., Miller, Price, Okun, Montijo, & Bowers, 2009; Ragland, et al., 2002). The degree of difficulty associated with an *n*-back task (i.e., load factor) can be manipulated by increasing the number of steps back into the sequence the participant must keep in memory, in order to determine a matching stimulus. Thus, in the commonly administered 2-back task, the participant must keep in memory information presented at two trials back, in order to determine whether an incoming stimulus is consistent with the target or not.

Despite the frequency with which *n*-back tasks are used in research, only a small number of psychometric evaluations of *n*-back test scores appear to have been published. Furthermore, the sample sizes upon which the psychometric analyses were performed tend to be relatively small (N = < 100) and non-representative of the general population (i.e., university students). For example, Hockey and Geffen (2004) administered a visuospatial *n*-back task across four load conditions: 0, 1, 2, and 3. The *n*-back task was administered a second time 1-week later to estimate test-retest reliability. For the purposes of evaluating convergent validity, Hockey and Geffen (2004) also administered five subtests from the Multidimensional Aptitude Battery (Jackson, 1998). The sample consisted of 70 university students (FSIQ = 120.12; SD = 12.03). Both reaction time and accuracy scores were derived from the *n*-back scores. The test-retest reliabilities for the 0-back, 1-back, 2-back, and 3-back reaction time scores corresponded to .86, .79, .69, and .81, respectively, which would suggest moderate stability in performance. By contrast, the test-retest reliabilities for the corresponding accuracy scores corresponded to .52, .49, .54, and .73, which suggests an unacceptably low level of performance consistency across time. Unfortunately, the sample size was too small to interpret any numerical differences meaningfully. Also, internal consistency reliabilities were not calculated for any of the *n*-back scores.

Both reaction time and accuracy did evidence theoretically congruent correlations with the Multidimensional Aptitude Battery, however. For example, 2-back reaction time and 2-back accuracy correlated -.30 and .29 with FSIQ, respectively (Geffen, 2004). However, perhaps somewhat disturbingly, no significant correlations were observed between *n*-back reaction time and *n*-back accuracy, which suggests that they are measuring different dimensions of cognitive functioning. The issue is important to consider, as some who investigation working memory capacity believe that complex span tasks yield two dependent variables relevant to working memory capacity: accuracy and speed (e.g., Logan, 2004). However, as discussed further below, processing speed is a dimension of cognitive ability that is considered separate from other types of cognitive abilities.

Other commonly used tests of working memory that fall into Conway et al.'s (2005) category of complex span include reading span (Daneman & Carpenter, 1980), operation span, and counting span. These tasks have in common a memory element and an intervening processing element. For example, for the typical reading span task, a participant is required to read aloud a series of sentences (usually nonsensical). The sentences are read aloud and overheard by the test administrator to ensure attention is directed to a cognitive task: reading. Each sentence is 13 to 16 words in length. Crucially, the participant must also recall the last word associated with each sentence that is read aloud. As per other memory span tests, testing is terminated once a discontinue rule is reached (e.g., all three trials within an item; Daneman & Carpenter, 1980). The maximum number of sentences administered within an item is usually around six, which implies that the maximum score possible is six. Redick et al., (2012) provided normative information for several complex span tasks (N = 5,537).

Memory span, particularly complex memory span, is an active area of test development. Several batteries have been published for public use and are available free of charge (e.g., Hicks, Foster, Engle, in press; Lewandowsky, Oberauer, Yang, & Ecker, 2010; Oswald, McAbee, Redick, & Hambrick, 2015). Although all of these test batteries are computer administered, none have yet taken advantage of the benefits of computer adaptive testing. Theoretically, a complete memory span battery that was computer adaptive in nature would reduce testing time and enhance the reliability associated with test scores, as participants would complete more items close to their maximal capacity (Gershon, 2005). Memory span tests are especially attractive candidates for adaptive testing, because there is no need to first calibrate the item difficulties in a large, population representation sample.

Distinguishing Difficulty from Complexity

In pure psychometric terms, difficulty refers to the proportion of individuals who can solve a test item successfully (Raykov & Marcoulides, 2011). However, it is useful to distinguish difficulty from complexity. In the context of cognitive ability testing, complexity refers to the number of distinct cognitive processes recruited during the execution of a task. Theoretically, relatively more complex tasks require the recruitment of a relatively larger number of cognitive processes.

For example, there is an appreciable amount of empirical evidence to suggest that the serial recall of objects in a forward format is, on the whole, easier than the serial recall of objects in backward format. Based on a combination of several normative samples, Gignac (2015a) estimated that a typical healthy adult can recall 6.56 digits forward. By contrast, a typical healthy adult can only recall 4.88 digits backward. It would be simplistic to suggest that digit span forward is less difficult than digit span backward for several reasons. In particular, a digit span forward test can be created to be much more difficult than a digit span backward test. For example, if all of the digit span forward items included trials with a minimum of 9 digits, only a small percentage of participants would be expected to complete any of them successfully (≈ 10%; Wechsler, 2008). Thus, the digit span forward test would be considered difficult (i.e., even more difficult than a typical digit span backward test which included series of trials with a range of digits spaning from 2 to 8).

Despite the increase in test difficulty associated with the digit span forward test described above, it would be arguably inappropriate to suggest that it was more complex than a typical digit span backward test. The reason is that the difficult digit span forward (DSF) test would be expected to recruit fewer process to complete, in comparison to a moderately difficult digit span backward (DSB) test. In fact, there is some empirical research to suggest that a typical digit span backward test recruits additional (unique) visuospatial processes to complete. Colom, Jung, and Haier (2007) found that the execution of DSB recruited nine right-hemispheric parietal and temporal regions of the brain, whereas DSF recruited only four right-hemispheric parietal and temporal regions. The larger number of areas recruited during the execution of DSB suggests that it is a more complex task to complete, rather than simply more difficult. It is suggested here that a DSF item with 5 digits would recruit the same number of areas of the brain to execute as a DSF item with 9 digits: thus, the two items would be unequally difficult, but equally complex from an intelligence perspective.

A relatively complex test may be expected to be associated with a complex pattern of cross-loadings within a factor analysis, as it involves the recruitment of a number of different group-level processes. However, the existing factor analytic research does not support such a position. Instead, relatively complex tests tend to load more substantially upon the general factor of intelligence, as discussed further below.

Long-Term Memory and Crystallized Intelligence

The definition of long-term memory, as distinguished from short-term memory, is to some degree arbitrary, as the key distinguishing psychometric feature is time: the time elapsed between the presentation of the stimuli and recall. In contrast to short-term memory and working memory, intelligence researchers have focussed much less on the measurement of long-term memory. For example, the WAIS-IV does not include subtests relevant to the measurement of long-term memory. However, the CHC model of intelligence does clearly distinguished long-term memory from short-term memory (Glr; McGrew, 2009). In contrast to most other batteries, the Woodcock-Johnson-IV (Schrank, McGrew, & Mather, 2014) includes two subtests of long-term memory, Story Recall and Audio Visual Learning. In Story Recall, for example, the examinee listens to a story and then relates the story back to the psychometrist. The Kaufman Adult and Adolescent Intelligence Test (Kaufman & Kaufman, 1993) includes a similar subtest named Auditory Comprehension; however, curiously, it is considered a measure of crystallized intelligence.

The lack of consistency in the categorization of long-term memory tests suggests a lack of clarity in the literature about the nature of Glr. In relatively abstract terms, crystallized intelligence has been suggested to represent the degree to which a participant has internalized his/her culture. In more operational terms, crystallized intelligence represents the amount of general, or “every day”, knowledge a person has accumulated in his/her life and is able to express at the time of testing. Typically, researchers contextualise crystallized intelligence as linguistic in nature (e.g., McGrew, 2009), however, there would be expected to be visual-spatial elements of crystallized knowledge, as well. For example, a famous faces test (e.g., Lander & Poyarekar, 2015) would be expected to tap into crystallized intelligence, as would a visual identification knowledge test of the planets and moons within our solar system.

Prototypical measures of crystallized intelligence include vocabulary and knowledge of worldly facts. A typical vocabulary test involves the presentation of a word to a participant who is then required to either articulate the definition of the word (Wechsler, 2008), or identify the correct definition amongst a series of alternatives in a multiple choice format (e.g., MAB). For most tests of vocabulary, the participant is only required to express or identify a synonym of the target word, in order to achieve full marks. The extent to which such a limited requirement attenuates the validity of test scores is an interesting question.

Measures of crystallized intelligence have been argued to be heavily influenced by educational opportunities and experiences (Brown, 1994). For this reason, measures of crystallized intelligence have been criticized as less valid indicators of intellectual functioning, because they represent “learned intelligence” (Ford, 2004, p. 18). At least partial support for a critical stance in relation to crystallized measures of intelligence can be obtained from the empirical literature. For example, cognitive ability subtests that measure vocabulary and knowledge of worldly facts have been found to be more substantially associated with years of education completed and age. Based on the WAIS normative sample, Birren and Morrison (1961) reported the correlation between Information and Vocabulary at .81, which suggested that 66% of their variance was shared. Additionally, substantial correlations between the Information and Vocabulary subtests and years of education completed were reported at .66 and .62, respectively. However, controlling for years of education completed and age, the correlation between Information and Vocabulary reduced only to .68, which suggested that 46% of the variance shared between Information and Vocabulary was independent of the effects of years of education completed and age. Thus, although years of education and age may possibly influence performance on crystallized subtests such as Information and Vocabulary, clearly, there is more to the story.

One reason crystallized measures of intelligence may be associated with a surprisingly substantial amount of validity, independent of the effects of education and age, is that the accumulation of a vocabulary and knowledge of worldly facts rests upon the ability to comprehend and reason with language (verbal comprehension), rather than simply store information in long-term memory, as suggested by some (e.g., Groth-Marnat, 1997). In fact, based on a confirmatory factor analytic investigation, Kan, Kievit, Dolan, and van der Maas (2011) found that crystallized intelligence and verbal comprehension were statistically indistinguishable after controlling for years of education completed. Thus, the unique variance shared between subtests such as Vocabulary and Information may represent a broader verbal comprehension construct.

With respect to memory span and crystallized intelligence, it is interesting to note that, based on the Woodcock-Johnson – Revised normative sample, Gignac (2016) reported a latent variable correlation of .56 between Glr and Gc. Thus, higher levels of long-term retrieval capacity were related to greater crystallized ability, as one might expect. Curiously, however, the latent variable correlation between short-term memory capacity (Gsm) and Gc was appreciably larger at .68. Thus, crystallized intellectual ability related more strongly with short-term memory processes, in comparison to longer-term memory processes. More research to help explain this counter-intuitive effect is encouraged.

Taken all of the above into consideration, the ability to accumulate knowledge does not appear to be a simple function of educational opportunity, age, and long-term memory capacity. Instead, the capacity to accumulate knowledge in everyday life appears to rest upon processes that are integral to intellectual functioning (e.g., reasoning). It is arguably for this reason that crystallized measures of intelligence have been observed to be associated with a substantial amount of predictive validity (Jensen, 2001). Additionally, it is worth noting that test scores derived from crystallized tests are substantially heritable. Based on a sample of 194 18-year-old twin pairs and test scores from the WAIS and Raven's, Rijsdijk and Vernon (2002) found that Information and Vocabulary were the two most genetically determined tests (76% and 72% heritability, respectively). By comparison, Raven's was associated with 64% heritability.

The effects of environmental influences on crystallized intelligence test scores are also interesting to consider. In particular, although the positive correlation between years of education and crystallized intelligence test scores is essentially undeniable, the direction of the effect is much less obvious. For example, Rijsdijk et al. (2002) failed to observe a statistically significant effect of shared environment on Information and Vocabulary test scores. Such a result suggests that the educational opportunities made available by the family have little or no impact on crystallized intelligence. In younger children, the shared environment effects on verbal ability test scores appears to be small but significant (Tambs, Sundet, & Magnus, 1986); furthermore, the small effect diminishes over time in favor of genetic effects (Rietveldt, Dolan, van Baal, & Boomsma, 2003). Arguably, an individual's cognitive ability genetic constitution has a developmentally increasing impact on the person's selected environment, which may impact the acquisition of knowledge over time.

If measures of crystallized intelligence, such as Vocabulary and Information, are essentially redundant with test scores from more pure measures of verbal comprehension, such as Similarities, Comprehension, and Miller Analogies (Miller, 1960), it is useful to raise the question of whether there are any benefits to the administration of tests more saturated by crystallized intelligence. Clearly, if a researcher is more interested in crystallized intelligence than verbal comprehension for theoretical reasons, then, yes, tests such as Vocabulary and Information would be expected to yield crucial information, particularly if they can be residualized of their verbal comprehension variance. In researcher contexts, the administration of crystallized measures may also prove useful to help specify a more robust verbal compression latent variable (i.e., with four indicators, rather than just two).

Progress in the measurement of crystallized intelligence has been especially stagnant, with few improvements implemented, or even suggested, over the years. I believe research in the area of language development has some interesting insights to offer psychometricians. For example, the distinction between vocabulary size (number of known words) and vocabulary depth (how well those words are known) appears to offer a more sophisticated approach to the conceptualization and measurement of crystallized intelligence (Schmitt, 2014). Currently, crystallized measures of intelligence do not distinguish these two types of abilities, despite the current evidence reported by Schmitt (2014) which suggests it is valuable to do so in the middle to upper-end of the distribution of ability.

Processing Speed and Chronometric Tests

Recall the abstract definition of intelligence adopted in this work: an entity's maximal capacity to achieve a novel goal successfully using perceptual-cognitive abilities. Although the definition does not include the word "quickly", it is the case that, today, intelligence measurement often includes individual differences in processing speed as an indicator of intelligence. Intelligence test developers do so for two reasons. First, it has been established that relatively intelligent individuals tend to process information relatively quickly. For example, the correlation between processing speed and crystallized intelligence, two ostensibly different constructs, inter-correlate positively at approximately $r = .49$ (Gignac, 2016). In fact, when tested on population representative samples, processing speed has been observed to correlate positively with all other measures of intelligence (Gignac, 2016). Thus, intelligence test developers view processing speed as a fundamental pillar to intelligence. The second reason cognitive ability tests may be developed with a time-limit is principally practical: to limit the time of test administration. There are substantial individual differences in the time taken to provide an answer to a question. Arguably, a non-negligible percentage of such variance is due to personality, rather than cognitive ability. Thus, by “forcing” participants to complete within a specified amount of time, the process of test administration can be contained or managed.

The downside of including a time-limit to the test administration instructions of a non-processing speed test, or the addition of bonus marks for quick completion, is that the test scores derived from such tests will be confounded by at least two sources of group-level factor variance: processing speed and the group-level factor of principal interest. Such test score group-level factor variance amalgamation can have important consequences: namely, inflated correlations between pure processing speed tests and theoretically non-processing speed tests. In an elegant study, Chuderski (2013) found that the correlation between fluid intelligence and working memory capacity increased as a function of the degree to which the measures both included a speeded element.

Researchers must make a number of important decisions when using reaction time (RT) data for analyses. A review of the literature suggests that there is little consistency, or even regard, for using methods that are considered more optimal than others. For example, should only RTs from accurate responses be used in the calculation of a participant's RTs? Arguably, only RTs associated with accurate responses should be used and reported, however, such a procedure is likely only defensible when the error response rates are low (≤ 5%; Jensen, 2005). When error rates are greater than 5%, researchers should estimate mean (or median) RTs for both the accurate and inaccurate responses, separately, and estimate their inter-association. If the correlation is large (> .85), then deriving participant RTs from a combination of the accurate and inaccurate trials will likely be defensible. However, if the correlation is not every large, the results should be examined and reported separately for both type of RT (i.e., accurate and inaccurate).

It is also alarming to observe the frequency with which reaction time and movement time (MT) are not measured separately in individual differences RT studies. Reaction time is also known as decision time: it represents the amount of central processing time required to make a decision. By contrast, movement time is the amount of time it takes to execute a peripheral motor response. Several researchers have argued that RT is the key construct of interest, not MT (e.g., Jensen, 1998). Jensen and Vernon (1986) reported a meta-analytically derived correlation of -.22 between RT (3-bits) and IQ, based on studies that used the Hick paradigm. Reaction time can be measured without the confound of movement time by employing a modified Hick paradigm (see Neubauder, 1991; Neubauer, Bauer, & Holler, 1992).

Jensen and Munro (1979) discovered that reaction time increases essentially monotonically ($r = 1.0$) with the amount of information that is processed. By contrast, movement time evidenced a much smaller effect (*r* = .54). Correspondingly, RT and MT were far from perfectly correlated (*r* = .37), which supports the notion that RT and MT are important dimensions of cognition/behavior to dissociate psychometrically. Jensen and Munro (1979) also found that both RT (*r* = -.39) and MT (*r* = -.43) correlated with fluid intelligence as measured by the Raven's. Carlson and Jensen (1982) replicated the result based on a small sample size of ninth graders (*N* = 20). Thus, it is important to distinguish MT and RT in reaction time studies, although few appear to do so.

How many trials are required to obtain an accurate RT is an important question to consider? Jensen and Munro (1979) used 30 trials and obtained international consistency reliabilities of .90 and .89 for RT and MT, respectively. Thus, 30 trials should be considered sufficient for the purposes of estimating an individual's RT, in most cases. However, many studies published in the area of cognition and individual differences appear to administer far more trials than is necessary to achieve a reliable estimate (> 50 trials). The amount of fatigue (and/or boredom) experienced by participants during such unnecessary testing is probably substantial, which would be expected to affect the validity of the RT scores adversely.

Another important decision researchers who measure reaction time must make is whether to calculate each individuals' mean RT or median RT as the indicator of typical (central tendency) processing speeds. Researchers can also usefully calculate individual differences in intra-individual variability (Jensen, 1992). Finally, measurement and analysis of the worst performance rule can also offer important insights (e.g., Rammsayer & Troche, 2016). In my view, an investigation that collects RT and/or inspection time data should seriously consider all four primary indicators of processing speed, as they appear to offer unique pieces of information.

A comprehensive treatment of the measurement of reaction time and inspection time can be found in Jensen (2006).

Reasoning

Individual differences in reasoning reflect the ability to make correct inferences from information (Burt, 1922; Lohman & Lakin, 2011). To experts and laypeople alike, reasoning is one of the primary characteristics of intelligent behavior (Sternberg, Conway, Ketron, & Bernstein, 1981). One of the central conceptual characteristics of reasoning involves moving beyond the information that is available (Bruner, 1957), which underscores the notion of novelty in the conceptualization and measurement of cognitive ability. Within the Cattell-Horn-Carroll model of intelligence, reasoning falls into the fluid intelligence (a.k.a., fluid reasoning) group-level factor (McGrew, 2009). Some work has suggested that fluid intelligence is the essence of general intellectual functioning (e.g., Gustafsson, 1984); however, such studies have been based on insufficiently large small sample sizes ($N < 250$), non-representative samples, and/or a collection of subtests that should be considered insufficiently broad to justify such a conclusion. Instead, it would be more justifiable to contend that fluid intelligence is the most substantial indicator of general intellectual functioning. For example, based on the WAIS-IV normative sample, Gignac (2014) found that fluid intelligence, measured by Figure Weights, Matrix Reasoning, and Block Design, related to general intelligence at .94; thus, 88% of the variance in fluid intelligence was accounted for by *g*.

Like most any other group-level cognitive ability, measures of fluid ability can be classified broadly as verbal and non-verbal in nature. Arguably, the non-verbal measures of reasoning have gained the most popularity. Perhaps the two most well-known measures of fluid intelligence include the Culture Fair Intelligence Test (CFIT; Cattell, 1963, 1973) and Raven's

Progressive Matrices (Raven, 2000). Both the CFIT and Raven's measure individual differences in the ability to identify a pattern amongst a series of figural images. Furthermore, the CFIT and Raven's are typically considered to be "culture reduced" cognitive ability tests, as they are less obviously affected by educational opportunities and experiences (e.g., Brouwers, van de Vijver, & van Hemert, 2009; Reeve, 2009). For this reason, test scores derived from fluid intelligence tests such as the CFIT and Raven's are viewed more favorably as indicators of intelligence.

An attractive feature associated with the CFIT is that it is composed of four subtests with slightly different stimulus presentations. Thus, test scores from the CFIT are not limited to a single approach to the measurement of Gf, a theoretical advantage over Raven's (Colom & Abad, 2007), although the four approaches are highly similar in nature. From a psychometric perspective, a less attractive feature associated with the four CFIT subtests is that they are only comprised of 10 to 14 items, consequently, the subtest internal consistency reliabilities are often observed to be rather low (< .60; e.g., Furlow, Armijo-Prewitt, Gangestad, & Thornhill, 1997; Gignac, Shankaralingam, Walker, & Kilpatrick, 2016). Additionally, the norms – which were never particularly impressive – are outdated, as the CFIT has not been updated since the second edition (i.e., 1973). There are also, arguably, an insufficient number of difficult items within each of the four subtests. Thus, first-year university students often achieve mean IQ levels > 120 (see Gignac et al., 2016), which is not consistent with the more up-to-date Wechsler scales (university student $M \approx 110$;Weyandt, Oster, Gudmundsdottir, B. G., DuPaul, G. J., & Anastopoulos, in press).

Additionally, the CFIT manual specifies a limit of 3 minutes for each of the four CFIT subtests, which, in practice, implies that a significant percentage of participants do not complete all of the items, or they do so under significant time pressure. Thus, in my view, CFIT test scores are appreciably confounded by processing speed (when the 3 minute time limit is imposed), although a purpose developed study does not appear to have yet been conducted to evaluate this contention.

Raven's progressive matrices exist in three forms: Raven's Coloured Progressive Matrices (for children; Raven, Raven, & Court, 1998); Raven's Progressive Matrices (general population; Raven, Raven, & Court, 2000); and the Advanced Progressive Matrices (top 25% of the population; Raven, Raven, & Court, 1998). Each of the three versions of Raven's consists of a single series of items, unlike the CFIT, which consists of four slightly different subtests. For example, the Standard Progressive Matrices consists of 60 items/diagrams and the Advanced Progressive Matrices consists of 36 items/diagrams. Each item/diagram is associated with eight alternative patterns, and the participant must choose one to complete the spatial pattern depicted in the item/diagram.

The length of time required to administer the full Raven's (≈ 45 minutes) has inspired work to develop short-forms, several of which have become rather popular. For example, Arthur and Day (1994) created a 12-item version of the APM which was reported to require a maximum of 20 minutes to administer. Although the Arthur and Day (1994) short-form evidenced reasonable internal consistency ($\alpha = .72$) reliability and concurrent validity (concurrent validity $r = .90$), Bors and Stokes (1998) noted that the first three items (items 1, 4, and 8) were rather easy to complete for university students, consequently, they recommended an alternative 12-item short-form which included only on item within the first 10 of the APM full-form (item 3). Bor and Stokes (1998) also recommended the administration of only two practice items. Additionally, the Bor and Stokes (1998) short-form was reported to be internally consistent ($\alpha = .73$) and possess a concurrent validity coefficient of .92 In my view, the sample sizes upon which these two studies were based (< 500) do not offer sufficient power to yield stable solutions for the purposes of selecting items with the most discriminatory power. Thus, some capitalization on chance likely influenced the results. Hamel and Schmittman (2006) administered the APM (all 36 items) to two groups: untimed ($N = 397$); and 20-minute time limit ($N = 51$). Hamel et al (2006) did not find a statistically significant difference in APM performance between the two groups. Additionally, they found that the 20-minute time limit did not affect the concurrent validity of the test scores, based on correlations between the APM and an overall intelligence score based on six cognitive ability tests (timed: $r = .55$; untimed: $r = .42$). Thus, imposing a 20-minute time limit on the APM does not appear to affect its validity, although it would be useful to replicate this investigation on another sample with a larger sample size in the timed group. Based on the results of Hamel et al., it may be speculated that an APM short-form based on only half the items (say, odd or even items) with a 10 to 12 minute testing time may also yield reasonably valid test scores, as well.

General Intelligence

General intelligence refers to the cognitive phenomenon that pervades a diverse collection of narrower cognitive abilities. The empirical basis of the general intelligence construct rests, in part, upon the positive manifold: the observation of positive correlations across a diverse collection of cognitive ability tests (Burt, 1939; Jensen, 1998). A substantial amount of empirical research has demonstrated that general intelligence, or g, is the best cognitive predictor of a number of consequential events in life (e.g., years of education completed, job performance, life expectancy; Gottfredson, 1997).

In comparison to the narrower cognitive ability dimensions discussed in this chapter, g has attracted a substantial amount of criticism. For example, some have contended that g is a statistical artifact (Bartlett, 1937; Thomson, 1916), an epiphenomenon of human development (van der Maas et al., 2006), and, finally, a pseudo-scientific basis to justify racist ideologies (Gould, 1981).

As general intelligence is a cognitive phenomenon that pervades a number of narrower cognitive abilities, it stands to reason that general intelligence cannot be measured with a single test. Jensen (1998) recommended that general intelligence be measured with a minimum of nine tests representative of at least three group-level ability dimensions (e.g., crystallized intelligence, fluid intelligence, processing speed, memory span). Researchers will likely rarely be led astray following such advice, providing the individual subtests are associated with respectable psychometric properties (reliability and validity). I discuss further below the prospects associated with measuring general intelligence with fewer subtests.

Several researchers appear to have difficulties with the general factor of intelligence, as it has been argued to be atheoretical and inexplicable from a cognitive process perspective (e.g., Conway & Kovacs, 2013). Such an argument is either unmerited or made unjustly, in my view. Consider, for example, that individual differences in simple verbal span correlate positively with complex spatial span (Kane et al., 2004). What cognitive process has been established definitively to account for such a positive correlation? Despite the absence of any such definitive account in the literature, the observation of a positive, but imperfect, correlation between simple verbal span and complex spatial span has not resulted in contentions that the construct of memory span is unjustifiable, or that observed indicators of memory span should not be used to form a latent variable. Arguably, there is a shared cognitive process that helps bind these different memory span dimensions together, in addition to those cognitive ability dimensions that are even more theoretically distinct, but also positively inter-correlated (i.e., the positive manifold).

Jensen (1998) claimed that the substantial association between working memory capacity and fluid intelligence may be mediated by individual differences in processing speed. Conway, Kane, and Engle (1999) countered by presenting empirical evidence which suggested that individual differences in controlled attention, not processing speed, mediated the association between WMC and *g*. Perhaps most compellingly, Conway et al. (1999) noted that variability in individual reaction times across trials, and/or the worst performance for a trial, were better predictors of fluid intelligence than individual mean (or median) reaction times (see Coyle, 2003; Schmiedek, Oberauer, Wilhelm, Süß, & Wittmann, 2007). Thus, deficits in sustained (or controlled) attention, as represented by variability in processing speed, reflects a fundamental attribute of intellectual functioning. Arguably, the same phenomenon may be used to help explain the positive association between all types of cognitive abilities, as it is difficult to conceive of cognitive ability tests which do not require controlled attention to performance successfully. In my view, the construct of sustained/controlled attention is simply a more contemporary and sophisticated description of Spearman's (1923) "mental energy" theory of *g*.

It should be noted that it is highly unlikely that any single test of controlled attention will ever be developed to represent fully the general factor of intelligence. Thus, the failure to observe large correlations between putative tests of controlled attention and fluid intelligence (or g) should not be viewed as evidence against the notion that controlled attention may be the core underlying processing mediating g. Instead, the capacity to exhibit maximal cognitive performance consistently across a diverse number of cognitive ability tests, over a non-negligible amount of time (> 30-60 minutes), is arguably the most valid approach to the measurement of controlled attention. It is also plausible to suggest that some test scores may reflect the capacity to sustain attention in a controlled manner over a period of many years (e.g., crystallized intelligence).

Another reason a single measure of controlled attention will unlikely ever be developed to represent general intelligence is that complexity plays an arguably important role in general intellectual functioning (Larson, Merritt, & Williams, 1988; Rammsayer & Troche, 2016; Stankov, 2000). As discussed above, DSB is a better indicator of general intellectual functioning than DSF, because DSB appears to draw upon a more complex network of cognitive resources than DSF (Colom, Jung, & Haier, 2007). Thus, in that sense, DSB is more complex. By contrast, a task such as the Stroop Test (Stroop, 1935), commonly described as a measure of controlled attention (West, 2004), is a relatively uncomplicated task, as it requires attention principally to a single dimension, while ignoring another. Correspondingly, Stroop's *g* loading is not particularly large (≈ .45; Burns, Nettelbeck, & McPherson, 2009). This view accords well with the process overlap theory of *g* described by Kovacs and Conway (2016). I would hypothesize that there are natural individual differences in the brain's capacity to recruit and utilize multiple processes simultaneously to execute cognitive tasks.

Brief Measurement of General Intelligence

A survey of the literature suggests that researchers who are interested in intelligence, but whose project parameters limit the amount of time available to measure intelligence, typically employ one of three strategies: (1) administer a single measure of fluid intelligence, such as the RAPM or the CFIT; (2) administer a specifically developed brief intelligence test, such as the WASI or the Kaufman Brief Intelligence Test (KBIT; Kaufman & Kaufman, 2004); or (3) administer a short-form (typically, two to seven subtests) derived from a comprehensive intelligence battery. As will be contended in this section of the chapter, none of the three alternatives may be considered particularly attractive, as most options within each of the three categories nonetheless require in excess of 30 minutes to administer. Furthermore, the reported validity associated with intelligence test scores derived from short administration times (< 15 minutes) is limited.

With respect to the first strategy, the RAPM consists of 36 items and can take between 40 and 60 minutes to administer (Arthur & Day, 1994). The CFIT (Form III) consists of four subtests and a total of 46 figural items (Cattell, 1973). Although the total amount of time the participant spends completing the CFIT items is 12.5 minutes, the total administration time is closer to 30 minutes (Motta & Joseph, 2000). Arguably, a testing time of between 30 and 60 minutes for a single measure is a substantial commitment, particularly considering that many projects may have allocated 45 to 60 minutes of total testing time.

Additionally, although some researchers appear to equate scores from the RAPM and/or the CFIT as pure indicators of *g* (e.g., Arthur & Day, 1994; Kunda, McGreggor, & Goel, 2009; Raven & Raven, 2003), *g* is considered more defensibly to be defined by a diversity of cognitive capacities (Gignac, 2014; Jensen, 1998; Oberauer, Schulze, Wilhelm, & Süß, 2005). There is ample empirical evidence which supports the position that Raven's is not valid representation of *g*. Consider, for example, that Wechsler (1997) reported a correlation of only .64 between the Standard Progressive Matrices (SPM) and WAIS-III FSIQ scores. Additionally, Frearson, Barrett, and Eysenck (1988) reported a correlation of .71 between the RAPM and WAIS-R FSIQ scores. In a comprehensive investigation, Gignac (2015b) evaluated the *g* saturation associated with Raven's across four relatively large and relatively representative samples. Across the four samples, Raven's was found to share approximately 46% of its variance with *g* (mean *g* loading

= .68), which was not found to be especially noteworthy, in comparison to other well-regarded tests of intellectual functioning.

The second strategy consists of administering a specifically developed brief intelligence test. For example, the WASI (Wechsler, 1999, 2011) is composed of four subtests: Vocabulary, Similarities, Block Design, and Matrix Reasoning. Although these subtests are also found within the two most recent comprehensive Wechsler Adult Intelligence Scales (WAIS-III, Wechsler, 1997; WAIS-IV, Wechsler, 2008a), the number and precise nature of the items associated with each of the similarly named subtests is not the same across the abbreviated and comprehensive batteries (Homack & Reynolds, 2007). Another relatively commonly administered brief intelligence test is the Kaufman Brief Intelligence Test (KBIT; Kaufman & Kaufman, 2004). The KBIT consists of three subtests: Verbal Knowledge, Riddles, and Matrices. Brief intelligence tests were introduced to replace the myriad of short-forms that had been proposed over the years (Kaufman & Kaufman, 2001). Arguably, however, the second strategy of administering a brief intelligence test suffers from the same critical problem associated with administering single test such as the RAPM: time to administer. With respect to the KBIT, testing time can be expected to be up to 30 minutes to complete (Homack & Reynolds, 2007). Axelrod (2001) found that the four- and two-subtest versions of the WASI (Wechsler, 1999) took 34 and 17 minutes to administer, respectively. In practice, many research projects may not be able to accommodate the time required to administer tests such as the WASI and the KBIT, even though they are meaningfully briefer than their corresponding comprehensive batteries.

With respect to the third strategy, the possibility of developing a short-form derived from a comprehensive battery is an old idea. Wechsler (1944) suggested that if time did not permit the administration of the entire Wechsler-Bellevue scale (Wechsler, 1939), then one may consider administering a short-form which consisted of only the five verbal subtests. Over time, many proposed short-forms have accumulated in the literature. Frank (1983) referenced more than 30 proposed short-forms of the Wechsler scales, ranging in size from two to seven subtests. Based on the WAIS-R normative sample (Wechsler, 1981), Kaufman, Ishikuma, and Kaufman-Packer (1991) appear to have proposed the short-form with the least amount of administration time. Specifically, Kaufman et al. proposed Information and Picture Completion as an attractive short-form dyad with an average administration time of 12 minutes in total. Furthermore, Kaufman et al. reported that the Information-Picture Completion dyad was associated with a concurrent validity coefficient of .88. In the context of short-form evaluations, a concurrent validity coefficient represents the corrected correlation between the short-form composite scores and the corresponding full-scale composite scores (Tellegen & Briggs, 1967). Ringe, Saine, Lacritz, Hynan, and Cullum (2002) proposed a short-form based on a combination of Information and Matrix Reasoning which was associated with an average testing time of 14 minutes. Based on a sample of neuropsychological patients (N = 196), Ringe et al. found that the Information and Matrix Reasoning dyad was associated with a concurrent validity coefficient of approximately .90. Several other short-forms have been developed and investigated recently (e.g., Reid-Arndt, Allen, & Schopp, 2011; van Duijvenbode, Didden, van den Hazel, & Engels, 2016; Wagner, Pawlowski, Yates, Camey, & Trentini, 2010), however, no proposed short-form appears to have yet cracked the 10 minute mark.

Based on my research, the combination of Digit Span Backward, Coding, and Similarities would involve only approximately 8 minutes testing time (Axelrod, 2001).[1] Furthermore, based on my evaluation of the WAIS-IV normative sample correlation matrix with Tellegen and Briggs' (1967) concurrent validity formula, the combination of Digit Span Backward, Coding, and Similarities is associated with a validity coefficient of .87, which is greater than the minimally acceptable validity coefficient of .82 recommended by Donders and Axelrod (2002). Thus, approximately 76% of the variance FISQ scores could be accounted for by the combination of Digit Span Backward, Coding, and Similarities. Despite the ostensibly impressive concurrent validity and time administration estimates, some incredulity may be merited.

If general intelligence functioning is mediated substantially by individual differences in "mental energy" (i.e., the ability for sustained concentration and/or attentional control), a testing session that lasts 60 minutes would necessarily test a person's capacity better than a testing session that lasts only 10 minutes, all other things equal. As noted by others (e.g., Kaufman & Kaufman, 2001; Kaufman, Ishikuma, & Kaufman-Packer, 1991), the vast majority of the short-

[1] According to Axelrod (2001), Digit Span (both Forward and Backward) requires, on average, 3.6 minutes to administer. Thus, half of 3.6 = 1.8. Coding requires, on average, 3 minutes to administer. Finally, Similarities requires, on average, 3.6 minutes to administer. Thus, 1.8 + 3.0 + 3.6 = 8.4 minutes. In my experience, the administration time of Similarities can be reduced somewhat (about 30 to 45 seconds) by using a response sheet that includes the most common answers to the items. In practice, such a response sheet allows for the possibility of circling responses, rather than writing them, which saves time.

form IQ research has simply extrapolated the validity coefficient (correlation with complete battery FSIQ scores) based on the correlation matrix associated with the full administration of the battery, rather than an evaluation of the test scores associated with the administration of only those subtests included in the short-form.

In a rare exception, Thompson, Howard, and Anderson (1986) evaluated the validity of a WAIS-R two-subtest short-form (Vocabulary and Block Design) and a four-subtest short-form (Vocabulary, Arithmetic, Block Design, and Picture Arrangement) across two independent groups (*N* = 30 each). A third group consisted of participants who were administered the 11 subtests of the WAIS-R, according to the standard order recommended by the manual (also *N* = 30). In the two-subtest short-form condition, the two subtests were administered first, followed by the remaining nine WAIS-R subtests. In the four-subtest short-form group, the four subtests were administered first, followed by the remaining seven subtests. Thompson et al. (1986) found that the two-subtest short-form and the four-subtest short-form overestimated FSIQs by 4.93 and 2.47 IQ points, respectively. The effect was essentially replicated in a sample of 80 university students (Thompson & Plumridge, 1999). These results question, to some degree, the validity of short-forms published in the literature.

It is worthwhile pointing out that purpose developed brief measures of intelligence rely upon the administration of subtests that are the relatively longest to administer. For example, Vocabulary and Matrix Reasoning within the WASI each take approximately 15 minutes to administer, which implies a total testing time of approximately 30 minutes. The fact that brief batteries of intelligence are only relatively brief, not brief in an absolute sense, is likely not a coincidence. The prospect of achieving a respectable level of criterion-related validity, not just concurrent validity, with the administration of a combination of tests which involve less than 20 minutes administration time is perhaps unrealistic.

Such a hypothesis could be tested in a conventional predictive validity study, where the order of cognitive ability test administration was specified differently across two groups. For example, in the first group, the Matrix Reasoning subtest within the WAIS-IV could be administered first followed by the remaining subtests. In the second group, all of the same tests would be administered, except the Matrix Reasoning subtest would be administered last. If the Matrix Reasoning subtest were found to predict the criterion more substantially in the second group than the first group, then it would imply that testing time is an important characteristic in the assessment of general cognitive functioning; a characteristic that could not be substituted in the brief assessment of intelligence. Such an observation would also suggest that g may be at least partially mediated by individual differences in sustained concentration.

Processing speed is also important to consider in the context of evaluating brief measures of cognitive ability. For example, Baddeley’s (1968) 3-minute reasoning test incorporates a 3-minute time limit. Consequently, few participants complete all 64 items (*M* = 31.55, *SD* = 12.23; Batey, Chamorro-Premuzic & Furnham, 2009), which would imply that the test scores represent processing speed, perhaps even more so than reasoning (Vernon, Nador, & Kantor, 1985). Although processing speed is a genuine dimension of cognitive functioning, any general factor extracted from a series of tests associated with significant time limits will necessarily be oversampled by processing speed to the detriment of other cognitive ability dimensions. Thus, the general factor would not be considered sufficiently representative to be considered a good quality indicator of general intellectual functioning.

Summary and Conclusion

It is important to emphasize that there are many other dimensions of intellectual functioning that were not reviewed in this chapter. In fact, Newton and McGrew (2010) identified approximately 15 group-level factors of intelligence and approximately 100 facets of cognitive ability (associated with various level of validity, to-date). Of course, any treatment of the topic within the confines of a book chapter will necessarily be limited.[2] My impression of the literature is that knowledge relevant to our understanding of cognitive abilities has accumulated much more impressively over the last couple of decades, in comparison to advancements in the measurement of cognitive abilities. Improvements in measurement will necessarily translate into better quality substantive discoveries. Consequently, it would be unfortunate if the current sense of complacency surrounding the measurement of cognitive abilities were to continue for another several decades. It is hoped the contents of this chapter may serve to help encourage or inspire further work in the area.

[2] Not to mention the limitations of the author who was humbled during the preparation of this chapter, given the vast amount of literature I discovered was unfamiliar to me.

References

Anderson, J. R. (2005). *Cognitive psychology and its implications* (6th ed.). New York, NY: Worth.

Arthur, W., & Day, D. V. (1994). Development of a short form for the Raven Advanced Progressive Matrices test. *Educational and Psychological Measurement, 54*, 394-403.

Axelrod, B. N. (2001). Administration duration for the Wechsler Adult Intelligence Scale-III and Wechsler Memory Scale-III. *Archives of Clinical Neuropsychology*, *16*, 293-301.

Baddeley, A. D. (1968). A 3 min reasoning test based on grammatical transformation. *Psychonomic Science*, *10*, 341-342.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 8, pp. 47-89). New York, NY: Academic Press.

Bartlett, M. S. (1937). Note on the development of correlations among genetic components of ability. *Annals of Eugenics*, *7*, 299-302.

Batey, M., Chamorro-Premuzic, T., & Furnham, A. (2009). Intelligence and personality as predictors of divergent thinking: The role of general, fluid and crystallised intelligence. *Thinking Skills and Creativity*, *4*, 60-69.

Binet, A., & Simon, T. (1916). Application of the new methods to the diagnosis of the intellectual level among normal and subnormal children in institutions and in the primary schools. *L'Année Psychologique, 11*, 245-336. In H. H. Goddard (Ed.), *The development of intelligence in children (The Binet-Simon Scale)* (E. S. Kite, Trans.; pp. 91-181). Baltimore, MD: Williams & Wilkins. (Original work published 1905; Translation published 1916).

Birren, J. E., & Morrison, D. F. (1961). Analysis of WAIS subtests in relation to age and education. *Journal of Gerontology*, *16*, 363-368.

Boring, E. G. (1923). Intelligence as the tests test it. *New Republic*, *36*, 35-37.

Bors, D. A., & Stokes, T. L. (1998). Raven's Advanced Progressive Matrices: Norms for first-year university students and the development of a short form. *Educational and Psychological Measurement*, *58*, 382-398.

Brouwers, S. A., Van de Vijver, F. J., & Van Hemert, D. A. (2009). Variation in Raven's Progressive Matrices scores across time and place. *Learning and Individual Differences*, *19*, 330-338.

Brown, D. T. (1994). Review of the Kaufman Adolescent and Adult Intelligence Test (KAIT). *Journal of School Psychology*, *32*, 85-99.

Brown, G. D. A., & Hulme, C. (1996). Non-word repetition, STM, and word age-of-acquisition: A computational model. In S. E. Gathercole (Ed.), *Models of short-term memory* (pp. 129-148). Hove, England: Psychology Press.

Bruner, J. S. (1957). Going beyond the information given. In J. S. Bruner (Ed.), *Contemporary approaches to cognition: A symposium held at the University of Colorado* (pp. 41-69). Cambridge, MA: Harvard University Press.

Burt, C. (1939). The relations of educational abilities. *British Journal of Educational Psychology*, *9*, 45-71.

Burns, N. R., Nettelbeck, T., & McPherson, J. (2009). Attention and intelligence: A factor analytic study. *Journal of Individual Differences*, *30*, 44-57.

Burt, C. (1922). The development of reasoning in school children. *Journal of Experimental Pedagogy*, *5*, 68-77.

Cattell, R. B. (1963). *The IPAT culture fair intelligence scales 1, 2 and 3* (2nd ed.). Champaign, IL: Institute for Personality and Ability Test.

Cattell, R. B. (1973). *Technical supplement for the Culture Fair Intelligence Tests Scales 2 and 3*. Champaign, IL: Institute for Personality and Ability Testing.

Chase, W. G., & Ericsson, K. A. (1982). Skill and working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 16, pp. 1-58). San Diego, CA: Academic Press.

Charness, N., Krampe, R., & Mayr, U. (1996). The role of practice and coaching in entrepreneurial skill domains: an international comparison of life-span chess skill acquisition. In K. A. Ericsson (Ed.), *The road to excellence* (pp. 51-80). Mahwah, NJ: Lawrence Erlbaum.

Cantor, J., Engle, R. W., & Hamilton, G. (1991). Short-term memory, working memory, and verbal abilities: How do they relate? *Intelligence*, *15*, 229-246.

Carlson, J. S., & Jensen, C. M. (1982). Reaction time, movement time, and intelligence: A replication and extension. *Intelligence*, *6*, 265-274.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive psychology*, *4*, 55-81.

Chuderski, A. (2013). When are fluid intelligence and working memory isomorphic and when are they not? *Intelligence*, *41*, 244–262.

Colom, R., & Abad, F.J. (2007). Culture Fair Intelligence Test. In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics* (pp. 204-205). Thousand Oaks, CA: Sage.

Colom, R., Jung, R. E., & Haier, R. J. (2007). General intelligence and memory span: Evidence for a common neuroanatomic framework. *Cognitive Neuropsychology*, *24*, 867-878.

Conway, A. R., Kane, M. J., & Engle, R. W. (1999). Is Spearman’s G determined by speed or working memory capacity? *Psycoloquy*, *10*(074).

Conway, A. R., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*, 769-786.

Conway, A. R. A., & Kovacs, K. (2013). Individual differences in intelligence and working memory: A review of latent variable models. *Psychology of Learning and Motivation*, *58*, 233-270.

Conway, A. R. A., Getz, S. J., Macnamara, B., & Engel de Abreu, P. M. J. (2010). Working memory and fluid intelligence: A multi-mechanism view. In R. J. Sternberg & S. B. Kaufman (Eds.), *The Cambridge handbook of intelligence* (pp. 394-418). Cambridge, United Kingdom: Cambridge University Press.

Corsi, P. M. (1972). *Human memory and the medial temporal region of the brain*. Unpublished doctoral dissertation. McGill University.

Coyle, T. R. (2003). A review of the worst performance rule: Evidence, theory, and alternative hypotheses. *Intelligence*, *31*, 567-587.

Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd ed.). Oxford, UK: Harper.

Colom, R., Rebollo, I., Palacios, A., Juan-Espinosa, M., & Kyllonen, P.C. (2004). Working memory is (almost) perfectly predicted by g. *Intelligence*, *32*, 277-296.

Dang, C. P., Braeken, J., Ferrer, E., & Liu, C. (2012). Unitary or non-unitary nature of working memory? Evidence from its relation to general fluid and crystallized intelligence. *Intelligence*, *40*, 499-508.

Davidson, J. E., & Downing, C. L. (2000). Contemporary models of intelligence. In R. J. Sternberg (Ed.), *Handbook of intelligence* (pp. 34-52). New York, NY: Cambridge University Press

De Groot, A. D., & Gobet, F. (1996). *Perception and memory in chess: Heuristics of the professional eye*. Assen: Van Gorcum.

Dehn, M.J. (2008). *Working memory and academic learning: Assessment and intervention*. Hoboken, NJ: Wiley.

Donders, J., & Axelrod, B. N. (2002). Two-subtest estimations of WAIS–III factor index scores. *Psychological Assessment*, *14*, 360-364.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*, 363-406.

Eysenck, H. J. (1988). The concept of "intelligence": Useful or useless?. *Intelligence, 12*, 1-16.

Eysenck, H. J. (1986). Toward a new model of intelligence. *Personality and Individual Differences*, *7*, 731-736.

Ford, D. Y. (2004*). Intelligence testing and cultural diversity: Concerns, cautions, and considerations*. Storrs: University of Connecticut, The National Research Center on the Gifted and Talented.

Fougnie, D., Zughni, S., Godwin, D., & Marois, R. (2015). Working memory storage is intrinsically domain specific. *Journal of Experimental Psychology: General*, *144*, 30.

Frank, G. (1983). *The Wechsler enterprise: An assessment of the development, structure, and use of the Wechsler tests of intelligence*. Oxford: Pergamon Press.

Frearson, W., Barrett, P., & Eysenck, H. J. (1988). Intelligence, reaction time, and the effects of smoking. *Personality and Individual Differences*, *9*, 497-517.

Furlow, F. B., Armijo–Prewitt, T., Gangestad, S. W., & Thornhill, R. (1997). Fluctuating asymmetry and psychometric intelligence. *Proceedings of the Royal Society of London B: Biological Sciences*, *264*, 823-829.

Gathercole, S. E. (1999). Cognitive approaches to the development of short-term memory. *Trends in Cognitive Sciences*, *3*, 410-419.

Gathercole, S. E. (1996). *Models of short-term memory*. Hove, England: Psychology Press.

Gathercole, S. E. (2006). Nonword repetition and word learning: The nature of the relationship. *Applied Psycholinguistics*, *27*, 513-543.

Gershon, R. C. (2005). Computer adaptive testing. *Journal of Applied Measurement*, *6*, 109-127.

Gignac, G. E. (2014). Fluid intelligence shares closer to 60% of its variance with working memory capacity and is a better indicator of general intelligence. *Intelligence*, *47*, 122-133.

Gignac, G. E. (2015). The magical numbers 7 and 4 are resistant to the Flynn effect: No evidence for increases in forward or backward recall across 85 years of data. *Intelligence*, *48*, 85-95.

Gignac, G. E. (2015). Raven's is not a pure measure of general intelligence: Implications for g factor theory and the brief measurement of g. *Intelligence*, *52*, 71-79.

Gignac, G. E. (2016). Residual group-level factor associations: Possibly negative implications for the mutualism theory of general intelligence. *Intelligence*, *55*, 69-78.

Gignac, G. E., & Weiss, L. G. (2015). Digit Span is (mostly) related linearly to general intelligence: Every extra bit of span counts. *Psychological Assessment*, *27*, 1312-1323.

Gignac, G. E., Shankaralingam, M., Walker, K., & Kilpatrick, P. (2016). Short-term memory for faces relates to general intelligence moderately. *Intelligence*, *57*, 96-104.

Giofrè, D., Mammarella, I. C., & Cornoldi, C. (2013). The structure of working memory and how it relates to intelligence in children. *Intelligence*, *41*, 396-406.

Gobet, F., & Campitelli, G. (2007). The role of domain-specific practice, handedness, and starting age in chess. *Developmental Psychology*, *43*, 159-172.

Gonzalez, C., Thomas, R. P., & Vanyukov, P. (2005). The relationships between cognitive ability and dynamic decision making. *Intelligence*, *33*, 169-186.

Gottfredson, L. S. (1997). Why *g* matters: The complexity of everyday life. *Intelligence, 24*, 79-132.

Gould, S. J. (1981). *The mismeasure of man*. New York, NY: Norton.

Groth-Marnat, G. (1997). *Handbook of psychological assessment* (3rd ed.). New York, NY: John Wiley & Sons.

Gustafsson, J. E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence*, *8*, 179-203.

Hamel, R., & Schmittmann, V. D. (2006). The 20-minute version as a predictor of the Raven Advanced Progressive Matrices Test. *Educational and Psychological Measurement, 66*, 1039-1046.

Hicks, K. L., Foster, J. L., & Engle, R. W. (in press). Measuring working memory capacity on the web with the online working memory lab (the OWL). *Journal of Applied Research in Memory and Cognition*.

Hockey, A., & Geffen, G. (2004). The concurrent validity and test–retest reliability of a visuospatial working memory task. *Intelligence*, *32*, 591-605.

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

Jackson, D. N. (1998). *Multidimensional aptitude battery II*. London, ON: Research Psychologist Press

Jensen, A. R. (1990). Speed of information processing in a calculating prodigy. *Intelligence, 14*, 259-274.

Jensen, A. R. (1992). The importance of intraindividual variation in reaction time. *Personality and individual Differences*, *13*, 869-881.

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

Jensen, A. R. (2001). Vocabulary and general intelligence. *Behavioral and Brain Sciences*, *24*, 1109-1110.

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

Jensen, A. R., & Munro, E. (1979). Reaction time, movement time, and intelligence. *Intelligence*, *3*, 121-126.

Jensen, A. R., & Vernon, P. A. (1986). Jensen's reaction-time studies: A reply to Longstreth. *Intelligence*, *10*, 153-179.

Kan, K. J., Kievit, R. A., Dolan, C., & van der Maas, H. (2011). On the interpretation of the CHC factor Gc. *Intelligence*, *39*, 292-302.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, *133*, 189-217.

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

Kaufman, J. C., & Kaufman, A. S. (2001). Time for the changing of the guard: A farewell to short forms of intelligence tests. *Journal of Psychoeducational Assessment*, *19*, 245-267.

Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Brief Intelligence Test* (2nd ed.). Circle Pines, MN: AGS Publishing.

Kaufman, A. S., Ishikuma, T., & Kaufman-Packer, J. L. (1991). Amazingly short forms of the WAIS-R. *Journal of Psychoeducational Assessment*, *9*, 4-15.

Kovacs, K., & Conway, A. R. (2016). Process Overlap Theory: A unified account of the general factor of intelligence. *Psychological Inquiry*, *27*, 151-177.

Kunda, M., McGreggor, K., & Goel, A. K. (2009, October). Addressing the Raven's Progressive Matrices Test of" General" Intelligence. In *AAAI Fall Symposium: Multi-Representational Architectures for Human-Level Intelligence*.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence*, *14*, 389-433.

Lander, K., & Poyarekar, S. (2015). Famous face recognition, face matching, and extraversion. *The Quarterly Journal of Experimental Psychology*, *68*, 1769-1776.

Larson, G. E., Merritt, C. R., & Williams, S. E. (1988). Information processing and intelligence: Some implications of task complexity. *Intelligence*, *12*, 131-147.

Lewandowsky, S., & Thomas, J. L. (2009). Expertise: Acquisition, limitations, and control. In F. T. Durso (Ed.), *Reviews of human factors and ergonomics* (Vol. 5; pp. 140-165). Santa Monica, CA: Human Factors and Ergonomics Society.

Lewandowsky, S., Oberauer, K., Yang, L. X., & Ecker, U. K. (2010). A working memory test battery for MATLAB. *Behavior Research Methods*, *42*, 571-585.

Li, D., Christ, S. E., & Cowan, N. (2014). Domain-general and domain-specific functional networks in working memory. *NeuroImage*, *102*, 646-656.

Logan, G. D. (2004). Working memory, task switching, and executive control in the task span procedure. *Journal of Experimental Psychology: General*, *133*, 218-236.

Lohman, D. F. (1989). Human intelligence: An introduction to advances in theory and research. *Review of Educational Research*, *59*, 333-373.

Lohman, D. F., & Lakin, J. M. (2011). Reasoning and intelligence. In R. J. Sternberg & S. B. Kaufman (Eds.), *The Cambridge handbook of intelligence* (2nd ed.; pp. 419-441). New York, NY: Cambridge University Press.

Lykken, D. T. (2005). Mental energy. *Intelligence*, *33*, 331-335.

Maehler, C., & Schuchardt, K. (2009). Working memory functioning in children with learning disabilities: does intelligence make a difference?.*Journal of Intellectual Disability Research*, *53*, 3-10.

Martin, P. R., & Fernberger, S. W. (1929). Improvement in memory span. *The American Journal of Psychology*, *41*, 91-94.

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, *37*, 1-10.

Messick, S. (1981). Constructs and their vicissitudes in educational and psychological measurement. *Psychological Bulletin*, *89*, 575-588

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88*, 355-383.

Miller, W. S. (1960). *Technical manual for the Miller Analogies Test*. New York, NY: The Psychological Corporation.

Miller, K. M., Price, C. C., Okun, M. S., Montijo, H., & Bowers, D. (2009). Is the n-back task a valid neuropsychological measure for assessing working memory? *Archives of Clinical Neuropsychology*, *24*, 711-717.

Mosing, M. A., Madison, G., Pedersen, N. L., Kuja-Halkola, R., & Ullén, F. (2014). Practice does not make perfect: No causal effect of music practice on music ability. *Psychological Science*, *25*, 1795-1803.

Motta, R. W. & Joseph, J. M. (2000). Group intelligence tests. In G. Goldstein & M. Hersen (Eds.), Handbook of psychological assessment (3rd ed.; pp. 131-146). Oxford, UK : Elsevier Science Ltd.

Neisser, U. (1979). The concept of intelligence. *Intelligence*, *3*, 217-227.

Neubauer, A. C. (1991). Intelligence and RT: A modified Hick paradigm and a new RT paradigm. *Intelligence, 15*, 175-192.

Neubauer, A. C., Bauer, C., & Höller, G. (1992). Intelligence, attention, motivation and speed-accuracy trade-off in the Hick paradigm. *Personality and Individual Differences*, *13,* 1325-1332.

Newton, J. H., & McGrew, K. S. (2010). Introduction to the special issue: Current research in Cattell–Horn–Carroll–based assessment. *Psychology in the Schools*, *47*, 621-634.

Noël, M. P., Désert, M., Aubrun, A., & Seron, X. (2001). Involvement of short-term memory in complex mental calculation. *Memory and Cognition*, *29*, 34-42.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw Hill.

Oberauer, K., Schulze, R., Wilhelm, O., & Süß, H.M. (2005). Working memory and intelligence—Their correlation and their relation: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, *131*, 61-65.

Oberauer, K., Süß, H. M., Schulze, R., Wilhelm, O., & Wittmann, W. W. (2000). Working memory capacity: Facets of a cognitive ability construct. *Personality and Individual Differences*, *29*, 1017-1045.

Orton, R. E. (1986). Do tests measure those constructs that people interpret them to measure? *Educational Theory*, *36*, 233-240.

Oswald, F. L., McAbee, S. T., Redick, T. S., & Hambrick, D. Z. (2015). The development of a short domain-general measure of working memory capacity. *Behavior Research Methods*, *47*, 1343-1355.

Paul, R. H., Lawrence, J., Williams, L. M., Richard, C. C., Cooper, N., & Gordon, E. (2005). Preliminary validity of "integneuroTM": A new computerized battery of neurocognitive tests. *International Journal of Neuroscience*, *115*, 1549-1567.

Plomin, R., & Petrill, S. A. (1997). Genetics and intelligence: What's new? *Intelligence*, *24*, 53-77.

Plomin, R., Shakeshaft, N. G., McMillan, A., & Trzaskowski, M. (2014). Nature, nurture, and expertise. *Intelligence*, *45*, 46-59.

Priem, R. L., & Butler, J. E. 2001. Tautology in the resource based view and the implications of externally determined resource value: Further comments. *Academy of Management Review*, 26, 57–-66.

Raaheim, K., & Brun, W. (1985). Task novelty and intelligence. *Scandinavian Journal of Psychology*, *26*, 35-41.

Ragland, J. D., Turetsky, B. I., Gur, R. C., Gunning-Dixon, F., Turner, T., Schroeder, L., ... & Gur, R. E. (2002). Working memory for complex figures: An fMRI comparison of letter and fractal n-back tasks. *Neuropsychology*, *16*, 370-379.

Rammsayer, T. H., & Troche, S. J. (2016). Validity of the worst performance rule as a function of task complexity and psychometric g: On the crucial role of g saturation. *Journal of Intelligence*, *4*, 5.

Raven, J. (2000). The Raven's progressive matrices: change and stability over culture and time. *Cognitive Psychology*, *41*, 1-48.

Raven, J., & Raven, J. (2003). Raven progressive matrices. In R. S. McCallum (Ed.), *Handbook of nonverbal assessment* (pp. 223-237). New York, NY: Kluwer Academic/Plenum Publishers.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 2: The Coloured Progressive Matrices*. Oxford, England: Oxford Psychologists Press/San Antonio, TX: The Psychological Corporation.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Raven manual: Section 4. Advanced Progressive Matrices*. Oxford: Oxford Psychologist Press.

Raven, J., Raven, J. C., & Court, J. H. (2000). *Raven manual: Section 3. Standard Progressive Matrices*. Oxford: Oxford Psychologist Press.

Raykov, T., & Marcoulides, G. A. (2011). *An introduction to psychometric theory*. New York, NY: Routledge.

Reid-Arndt, S. A., Allen, B. J., & Schopp, L. (2011). Validation of WAIS-III four-subtest short forms in patients with traumatic brain injury. *Applied Neuropsychology*, *18*, 291-297.

Reeve, C. L. (2009). Expanding the g-nexus: Further evidence regarding the relations among national IQ, religiosity and national health outcomes. *Intelligence*, *37*, 495-505.

Ringe, W. K., Saine, K. C., Lacritz, L. H., Hynan, L. S., & Cullum, C. M. (2002). Dyadic short forms of the Wechsler Adult Intelligence Scale–III. *Assessment*, *9*, 254-260.

Redick, T. S., Broadway, J. M., Meier, M. E., Kuriakose, P. S., Unsworth, N., Kane, M. J., ... &

Kuriakose, P. S. Measuring working memory capacity with automated complex span tasks. *European Journal of Psychological Assessment*, *28*(3), 164-171.

Reynolds, C. R. (1997). Forward and backward memory span should not be combined for clinical analysis. *Archives of Clinical Neuropsychology, 12*, 29-40.

Richman, H. B., Staszewski, J. J., & Simon, H. A. (1995). Simulation of expert memory with EPAM IV. *Psychological Review, 102,* 305-330.

Rietveld, M. J. H., Dolan, C. V., Van Baal, G. C. M., & Boomsma, D. I. (2003). A twin study of differentiation of cognitive abilities in childhood. *Behavior Genetics*, *33,* 367-381.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Rudkin, S. J., Pearson, D. G., & Logie, R. H. (2007). Executive processes in visual and spatial working memory tasks. *The Quarterly Journal of Experimental Psychology*, *60*, 79-100.

Schmiedek, F., Oberauer, K., Wilhelm, O., Süß, H. M., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology: General*, *136,* 414-429.

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning*, *64,* 913-951.

Schrank, F. A., McGrew, K. S., & Mather, N. (2014). *Woodcock-Johnson IV Tests of Cognitive Abilities*. Rolling Meadows, IL: Riverside Publishing

Slaney, K. L., & Racine, T. P. (2013). What's in a name? Psychology's ever evasive construct. *New Ideas in Psychology*, *31*, 4-12.

Spearman, C. S. (1923). *The nature of 'intelligence' and the principles of cognition*. Oxford, England: Macmillan

Stankov, L. (2000). Complexity, metacognition, and fluid intelligence. *Intelligence*, *28*, 121-143.

Sternberg, R. J., Conway, B. E., Ketron, J. L., & Bernstein, M. (1981). People's conceptions of intelligence. *Journal of Personality and Social Psychology*, *41*, 37-55.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643-662.

Tambs, K., Sundet, J. M., & Magnus, P. (1986). Genetic and environmental contributions to the covariation between the Wechsler Adult Intelligence Scale (WAIS) sub-tests: A study of twins. *Behavior Genetics*, 16, 475-491.

Tellegen, A., & Briggs, P. F. (1967). Old wine in new skins: Grouping Wechsler subtests into new scales. *Journal of Consulting Psychology*, *31*, 499-506.

Thomson, G. H. (1916). A hierarchy without: general factor. *British Journal of Psychology, 8*, 271-281.

Thomson, G. A. (1919). On the cause of hierarchical order among correlation coefficients. *Proceedings of the Royal Society, A, 95,* 400-408.

Thompson, A., & Plumridge, S. (1999). Two-and four-subtest short forms of the WAIS–R: A comparative validity study with a normal sample. *Psychological Reports*, *84,* 371-380.

Thompson, A. P., Howard, D., & Anderson, J. (1986). Two-and four-subtest short forms of the WAIS-R: Validity in a psychiatric sample. *Canadian Journal of Behavioural Science, 18,* 287-293.

Tulving, E. (2007). Are there 256 different kinds of memory? In J.S. Nairne (Ed.), *The foundations of remembering: Essays in honor of Henry L. Roediger, III* (pp. 39-52). New York, NY: Psychology Press.

van der Maas, H. L., Kan, K. J., & Borsboom, D. (2014). Intelligence is what the intelligence test measures. Seriously. *Journal of Intelligence*, *2*, 12-15.

van Der Maas, H. L., Dolan, C. V., Grasman, R. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. (2006). A dynamical model of general intelligence: the positive manifold of intelligence by mutualism. *Psychological Review*, *113*, 842-861.

van Duijvenbode, N., Didden, R., van den Hazel, T., & Engels, R. C. (2016). Psychometric qualities of a tetrad WAIS-III short form for use in individuals with mild to borderline intellectual disability. *Developmental Neurorehabilitation*, *19*, 26-30.

Vernon, P. A., Nador, S., & Kantor, L. (1985). Reaction time and speed-of-processing: their relationship to timed and untimed measures of intelligence. *Intelligence*, *9*, 357-374.

Vicari, S., Bellucci, S., & Carlesimo, G. A. (2003). Visual and spatial working memory dissociation: Evidence from Williams syndrome. *Developmental Medicine and Child Neurology*, *45*, 269-273.

Wagner, F., Pawlowski, J., Yates, D. B., Camey, S. A., & Trentini, C. M. (2010). Viability of IQ estimates based on Vocabulary and Block Design WAIS-III subtests. *Psico-USF*, *15*, 215-224.

Watkins, M. J. (1977). The intricacy of memory span. *Memory and Cognition*, *5*, 529-534.

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

Wechsler, D. (1944). *The measurement of adult intelligence* (3rd ed.). Baltimore, MD: Williams & Wilkins.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale–Revised*. New York, NY: Psychological Corporation.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale—Third Edition.* San Antonio, TX: Pearson Assessment.

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

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale-Fourth Edition: Administration and Scoring Manual*. San Antonio, TX: Pearson Assessment.

West, R. (2004). The effects of aging on controlled attention and conflict processing in the Stroop task. *Journal of Cognitive Neuroscience*, *16*, 103-113.

Weyandt, L. L., Oster, D. R., Gudmundsdottir, B. G., DuPaul, G. J., & Anastopoulos, A. D. (2016). Neuropsychological Functioning in College Students With and Without ADHD. *Neuropsychology*.

Williamson, V. J., Baddeley, A. D., & Hitch, G. J. (2010). Musicians' and nonmusicians' short-term memory for verbal and musical sequences: Comparing phonological similarity and pitch proximity. *Memory & Cognition*, *38*, 163-175.

Williams, J. M. (1991). *Memory Assessment Scales professional manual*. Odessa, FL: Psychological Assessment Resources.

Woods, D. L., Kishiyama, M. M., Yund, E. W., Herron, T. J., Edwards, B., Poliva, O., ... & Reed, B. (2011). Improving digit span assessment of short-term verbal memory. *Journal of Clinical and Experimental Neuropsychology*, *33*, 101-111.