Spring 2017

The development and validation of an automatic-item generation measure of cognitive ability

Hines Scott

THE DEVELOPMENT AND VALIDATION OF AN AUTOMATIC-ITEM-GENERATION MEASURE OF COGNITIVE ABILITY

by

Scott Hines, M.A., B.A.

A Dissertation Presented in Partial Fulfillment

Of the Requirements for the Degree

Doctor of Philosophy

COLLEGE OF EDUCATION

LOUISIANA TECH UNIVERSITY

May 2017

LOUISIANA TECH UNIVERSITY

THE GRADUATE SCHOOL

FEBRUARY 22, 2017
Date

We hereby recommend that the dissertation prepared under our supervision by

Scott Hines

entitled

THE DEVELOPMENT AND VALIDATION OF AN AUTOMATIC-ITEM-GENERATION MEASURE OF COGNITIVE ABILITY

be accepted in partial fulfillment of the requirements for the Degree of

Doctor of Philosophy

Supervisor of Dissertation Research

Head of Department

Psychology and Behavioral Sciences
Department

Recommendation concurred in:

Advisory Committee

Approved:

Director of Graduate Studies

Dean of the College

Approved:

Dean of the Graduate School

ABSTRACT

Cognitive ability is perhaps the most studied individual difference available to researchers, being measured quickly and effectively while demonstrating a predictable influence on many life outcomes. Historically, the evolution of the psychometric study of cognitive abilities has pivoted on the development of new and better methodologies allowing for a more complete and efficient capture of intellect. For instance, recent advances in computer and Internet technology have largely replaced traditional pencil-and-paper methods, allowing for innovative item development and presentation. However, concerns regarding the potential adverse impact and test security of online measures of cognitive ability, particularly in unproctored situations, are well documented and have limited the use of such measures in organizational settings. Methods, such as the use of multiple test forms and computer adaptive testing coupled with item exposure algorithms, have addressed some test-security concerns. However, these methods require the costly and tedious development of extensive item pools. The burgeoning area of automatic item generation potentially addresses many of the test-security and item-development concerns through the creation of assessment items based solely on an item model and a computer algorithm. Moreover, once the elements that contribute to item difficulty are calibrated, the psychometric properties of the items are known, meaning that little to no human review of the items is required before their use. The purpose of the current study was to develop an experimental non-verbal measure of cognitive ability through automatic item generation, using an innovative item type. Using a sample of 333 adults, the results of the current analysis support the proposed cognitive model's ability to explain item difficulty. Likewise, the temporal stability and predictive validity of the experimental measure are supported. In doing so, the experimental measure answers some of the test-security and item-generation concerns that are associated with the development and administration of cognitive-ability measures in organizational settings.

APPROVAL FOR SCHOLARLY DISSEMINATION

The author grants to the Prescott Memorial Library of Louisiana Tech University the right to reproduce, by appropriate methods, upon request, any or all portions of this Dissertation. It is understood that "proper request" consists of the agreement, on the part of the requesting party, that said reproduction is for his personal use and that subsequent reproduction will not occur without written approval of the author of this Dissertation. Further, any portions of the Dissertation used in books, papers, and other works must be appropriately referenced to this Dissertation.

Finally, the author of this Dissertation reserves the right to publish freely, in the literature, at any time, any or all portions of this Dissertation.

Author

Date 2-22-2017

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

The measurement of cognitive ability has been heralded as one of the crowning achievements of the psychological sciences (Lamb, 1994). For an investment of an hour or less, psychologists can gain insights into an individual's functioning that may not be uncovered through long and costly observations (Nettlebeck & Wilson, 2005). It is the easiest, most reliable, and most valid individual difference available to psychologists and researchers, measured cheaply and quickly (Cohen & Swerdlik, 2009; Furnham, 2008; Schmidt & Hunter, 1998). Moreover, the scores obtained from broad measures of cognitive ability (e.g., I.Q.) conform to the terms commonly used in society to describe individuals as intelligent or smart (Herrnstein & Murray, 1994). As such, the terms *cognitive ability* and *general mental ability* (GMA) are often used synonymously with intelligence (Gottfredson, 2002).

However, change is a defining feature of the cognitive-abilities research. Throughout the history of psychometric investigations of cognitive abilities, researchers have embraced methodological advances leading to better and more efficient methods of understanding the nature of intellect. For instance, the development of sophisticated statistical procedures such as exploratory and confirmatory factor analysis allowed cognitive theorists to peer beyond the data and develop models that explain the nature of intellectual ability (Carroll, 1993; Cattell, 1971; Guilford, 1988; McGrew, 1997; Spearman, 1904; Thurstone, 1938). These advanced statistical procedures have also aided in the nullification or outright dismissal of competing theories of intelligence that fail to produce consistent or logical evidence concerning their validity (Carroll, 2003; Keith & Reynolds, 2010). Thus, methodological advances have aided cognitive abilities researchers in the pursuit of greater clarity with regard to what it means to be clever.

Change is also a constant in the measurement of cognitive abilities. For instance, early investigations of intellectual functioning focused primarily on measures of sensory ability as proxy measures of intellectual ability (Hergenhahn, 2009). However, once it was demonstrated that sensory abilities failed to explain real-world performance (e.g., academic achievement), attention was turned to the measurement of higher level mental processes and their practical benefits in differentiating the performance of individuals. Likewise, the circumstances of World War I dictated a paradigm shift in assessment administration. The result of this shift was the advent of group testing, allowing for the quick and efficient collection of vast amounts of information on a large number of individuals for whom personnel decisions could be made (Boake, 2002). As such, cognitive-abilities research and measurement can be seen as an evolving field marked by innovation resulting in more accurate and efficient measures of intellectual ability (Drasgow & Olson-Buchanan, 1999; Gierl & Haladyna, 2012; Parshall & Harmes, 2009).

Throughout much of the 20th century, paper-and-pencil-based measures of cognitive abilities were the dominant medium by which intelligence was tested. As it has in almost all other areas of society, the technological revolution has transformed our daily lives. Computers are now small, fast, and cheap, allowing much of society to embrace their use (Chernyshenko & Stark, 2015). However, advances in computer and Internet technology have opened a new universe of methodological considerations from which cognitive ability can be tapped (Naglieri et al., 2004). While early versions of computer-based assessments were little more than direct translations of paper-and-pencil measures to a computerized medium, the measurement of cognitive ability is no longer restricted to static statements and images (Barak & English, 2002; Bartram, 2006). Rather, a diverse array of innovative and dynamic auditory and visual items can be administered via computer, potentially tapping cognitive ability in ways that were previously impossible to achieve (Parshall & Harmes, 2009). Moreover, computerized assessments realize practical benefits such as standardized item administration and automatic scoring, thus reducing error and improving test reliability. Likewise, administering computerized assessments online allows for an immense pool of test takers to sit for the same measure from anywhere in the world and at a time of their choosing, reducing the costs associated with testing programs (Drasgow & Olson-Buchanan, 1999; Naglieri et al., 2004). Thus, many test developers have embraced the technological revolution as more and more tests are being developed that can exploit the advantages afforded by computers and the Internet.

Despite the practical and measurement advantages offered by computer and online administration, problems in the areas of test construction and administration persist. For instance, large item pools are generally required as part of the test-development process. This problem is compounded when multiple forms of the same measure or advanced item-presentation methods (e.g., computer-adaptive testing) are used, necessitating an even larger number of items (Drasgow, Nye, Guo, & Tay, 2009). However, not all items that are created are useable. Despite the need for quality items and despite care taken to generate items that tap the construct of interest, many items must be removed at the item-analysis phase due to insufficient psychometric characteristics (Geerlings, Glas, & van der Linden, 2011; Wainer, 2002). This problem is particularly relevant to human item writers who often fail to construct items that conform to the construct of interest or at a desirable level of difficulty, further limiting the number of usable items (Hornke & Habon, 1986). Moreover, some cognitive researchers have questioned the validity of the results obtained from measures of intelligence administered in unproctored environments (Naglieri et al., 2004). The administration of measures in uncontrolled environments introduces a host of test-security threats (e.g., cheating) that distort test taker scores in ways that are difficult to detect. Since these distortions are not systematic, the validity of a measure is often greatly reduced due to the lessened predictive power it possesses (Foster, 2010). Thus, although technological advances have afforded greater options in how and where tests are administered, persistent issues remain that stunt researchers' ability to obtain convenient and accurate results.

The burgeoning arena of automatic item generation (AIG) seeks to address the concerns raised through the generation of a vast number of unique items strictly through an algorithm (Gierl, Ball, Vele, & Lai, 2015). Using an item model, the structural elements that relate to item difficulty are identified and manipulated, producing an array of items with known psychometric characteristics. Thus, little or no human review of the items is required before their administration (Doebler & Holling, 2015). Moreover, recent advances in AIG methodology allow for the creation of items directly from the calibrated structural elements, thereby addressing several issues associated with traditional test construction and administration (Geerlings, van der Linden, & Glas, 2012).

Despite the advantages posed through its use, limited research exists concerning the construction and subsequent validation of cognitive ability measures developed using AIG methodology (Gierl & Lai, 2012). Many researchers studying AIG measures have only examined the construct validation of the items, ignoring the predictive validity of these measures. Furthermore, the capability to create dynamically generated and presented items on-the-fly though AIG methodology has received little attention (Geerlings et al., 2012).

The purpose of the current research is to build on the existing AIG methodological framework through the construction and validation of an on-the-fly measure of cognitive ability that is generated at the time of item presentation. As such, this measure will not draw from a preexisting pool of items. Rather, the current measure will create items dynamically through predefined computer algorithms. The benefits of such a measure will address many of the issues that surround current test development. First, such a measure will be capable of generating a vast number of items through the use of an algorithm applied to an item model, producing items with known psychometric characteristics. As such, once calibrated, thousands of unique items of varying difficulty can be generated, without the need of human intervention. Second, many issues of test security will be addressed as each test taker will be administered different items. Although different items will comprise the measure for each test taker, the items will be calibrated such that the measure has identical construct adherence and psychometric properties. Third, once created, the criterion related validity of the measure will be assessed by examining the relationship that the experimental AIG measure shares with established indicators of cognitive ability. Thus, the experimental AIG measure created through this research is expected to advance the field's understanding of AIG item development and its relationship to other measures of cognitive ability.

Cognitive Ability

Although cognitive ability is one of the most studied individual differences in all of psychology (Gottfredson, 2002), reaching definitional agreement has proven problematic. In general terms, cognitive ability can be conceptualized as the basic mental capacity to reason, plan, solve problems, think abstractly, understand sophisticated and complex ideas, and acquire new information quickly and efficiently (Gottfredson, 2004). Similarly, Neisser (1967) defines intelligence as the "ability to understand complex ideas, to adapt effectively to the environment, to learn from experience and to engage in various forms of reasoning to overcome obstacles by taking thought" (p.7). Despite these seemingly straightforward descriptions of intelligence, substantial disagreement remains among cognitive ability theorists regarding the number of facets that are considered essential and how they should be arranged (Carroll, 1993; Cattell, 1971; Guilford, 1988; McGrew, 1997; Spearman, 1904; Sternberg, 1999). As quipped by Ackerman, Beier, and Boyle (2005), "there are as many intelligence theories as there are intelligence theorists..." (p. 31).

From a practical standpoint, people are readily able to recognize intelligence in others. Sternberg, Conway, Ketron, and Bernstein (1981) attempted to uncover this implicit conceptualization of intelligence by asking laypersons going about their daily lives in places such as grocery store parking lots to describe the behaviors associated with various portrayals of intelligence. The researchers then asked cognitive ability experts (i.e., psychologists) the same question. After analyzing the statements produced, the researchers found that although the two groups differed in their academic familiarity with the construct, both groups produced a pattern of relatively consistent descriptive terms of intellectual ability. For example, the attributes most associated with prototypical intellectual ability included problem solving, reasoning, and open-mindedness. Conversely, the hallmarks of unintelligence are characterized by personality trait-like behaviors (Costa & McCrae, 1992) including a lack of curiosity and a lack of tolerance of the views held by others. However, despite the consistency of responses obtained by the researchers, the variety of the descriptors of prototypical intellectual ability is indicative of the difficulty cognitive theorists have had in reaching definitional agreement.

Factor Analytic Theories

Although obtaining definitive agreement on a definition of intelligence has been elusive (Sternberg & Detterman, 1986), the emergence of several theories of intelligence can be traced to an important development in the field of statistics. In the early 1900s, Charles Spearman (1904) developed a primitive form of modern factor analysis allowing researchers to clarify the latent relationships shared by specific variables or phenomena. The purpose of factor analysis is to reduce and represent the number of observed variables into a smaller number of underlying hypothetical variables or "factors" (Anastasi & Urbina, 1997). The interpretation and measurement of each factor is dependent on a determination of the observed variables that make up the factor. Thus, by examining the interrelationships shown from factor analyzing measures of intellect, theorists are able to construct models of cognitive abilities that account for the results obtained.

Importantly, the latent structure of the factors that emerges from factor analysis is often open to interpretation. The choices made in conducting factor analysis (e.g., rotations, eigenvalue and factor loading cutoffs) complicate the convergence of interpretations that are made (DeVellis, 2012). Moreover, the labels that are applied to the factor(s) that emerge are dependent on the researcher's ability to subjectively determine the content and associated psychological processes of the measures that load most heavily on a particular factor. Thus, the interpretation of factors can be viewed as an art grounded in empirical data.

As stated by Humphreys (1962), "test behavior can almost endlessly be made more specific...factors can almost endlessly be fractionated or splintered" (p. 475). Thus, competing theories of intelligence have emerged stipulating a variety of structures and a diverse set of factors that make up cognitive functioning. However, as pointed out by Vernon (1950), only the factors that are "shown to have significant practical value in daily life are worth incorporating in the picture" (p. 25).

Two-Factor Theory. Perhaps the most famous and influential of the cognitive theorists is Charles Spearman. Noting that students who performed well on one measure of intelligence tended to perform well on other cognitive measures, Spearman (1904) used his factor analytic technique to identify the commonalities across performance across measures. Based on his findings, Spearman developed his theory of general intelligence, tapped by all measures of cognitive ability. Known as the Two-Factor theory of Intelligence, Spearman found that two factors or forms of intelligence emerged from the data: a general factor (*g*) and a test specific (*s*) factor. According to Spearman (1927), *g* is an innate general mental ability that contributes to all cognitive processes. This *g* factor is a "general fund of mental energy" (Spearman, 1914, p. 103) that explains why an individual's score on any given measure of cognitive ability is correlated with the scores obtained from other measures of cognitive ability. Conversely, *s*-factors explain why someone may obtain higher or lower scores on any given intellectual measure, but not performance across measures or task-domains. That is, specific factors, along with error, explain why performance on different cognitive measures is less than perfectly correlated. As such, *s*-factors do not add to the prediction of additional variance in cognitive ability because they operate only within specific measures of intelligence (Anastasi & Urbina, 1997). Thus, *g* and *s* are differentiated in that *g* is responsible for an individual's performance across all measures of cognitive ability, while *s*-factors are restricted to performance on independent measures of mental abilities. As such, *g* is thought to explain intellectual test performance (Jensen, 1998), conforming to what people describe as intelligence and leading most psychologists to adopt it as their operational definition of intelligence (Gottfredson, 2002). Thus, the importance of *g* to intellectual ability cannot be overstated as indicated by Ree and Earles (1993), "*g* is to psychology what carbon is to chemistry" (p.11).

Evidence of Spearman's theory is provided by the positive correlation observed across measures of cognitive ability. The *g* factor emerges regardless of whether mental test batteries are administered to different ages, sexes, races, and national groups and subsequently factor analyzed (Jensen, 1998). As such, although mental tests are designed to measure specific areas of cognitive functioning (e.g., verbal, spatial, and quantitative ability), individuals who perform well in one area, also tend to perform well on the others (Gardner, 1999), a phenomenon that Spearman termed "indifference of the indicator" (Spearman, 1927).

The core of cognitive ability research rests on this *positive manifold* (van der Maas, Kan, & Borsboom, 2014), the observation that the subtests of all intelligence tests ranging from academic measures to measures of social intelligence are positively correlated. *g* refers to a latent variable that results from the intercorrelation of several measures of cognitive ability (Spearman, 1927). Tests that correlated well with other measures of intelligence are indicative of higher levels of *g*-saturation. As such, *g*-saturation indicates the degree that a measure is tapping the general fund of mental energy. In contrast, cognitive tests that demonstrate a lesser relationship to other measures are thought to tap *s* factors such as residual variance due to test-specific abilities or otherwise contain error (e.g., unreliability). Therefore, higher levels of *g*-saturation are considered better predictors of intelligence. As such, Spearman suggests that a single highly *g*-saturated test be substituted for heterogeneous collections of tasks and items found in measures of intelligence (Spearman, 1927).

In order to select a measure that best approximated *g*, Spearman (1927) described *g* as the ability to extrapolate principles from one's experience and observations, best measured by abstract reasoning problems in formal tests. According to Spearman (1938), the Penrose and Raven (1936), later named Raven's Progressive Matrices (Raven & Court, 1989) well represented the abilities associated with *g*. As such, the defining characteristics of tests that tap *g* are non-verbal assessments of spatial or inductive reasoning.

Spearman noted that the Two-factor theory must be qualified, allowing for an intermediate class of factors that fall between *g* and *s*-factor. These intermediate factors, termed *group factors*, relate to some but not all intellectual tasks. Group factors are neither as universally broad as *g*, nor as specific as *s* (Anastasi & Urbina, 1997). Through the continued research of Spearman and his students, group factors such as mathematical, mechanical, and linguistic abilities were uncovered, laying the groundwork for future research and the development of more complex models of intellectual ability.

Primary Mental Abilities. On the heels of Spearman's work in identifying group factors, theories of intelligence moved from the existence of a single underlying mental ability to the identification of several abilities, and then to many. One such theory of multiple-intelligence was promoted by Louis Thurstone. Thurstone (1947) developed an advanced factor analytic technique allowing for the discovery of a multiple-factor structure of intelligence using orthogonal and oblique rotations, improving the interpretability of the data. Employing these techniques, Thurstone identified *g* as a second order factor subsuming narrower mental abilities.

Thurston (1938) concluded that intelligence could best be explained by seven *primary mental abilities*: Word Fluency, Verbal Comprehension, Number, Space, Perceptual Speed, Associative Memory, and Induction. Thus, in contrast to Spearman, Thurstone believed that cognitive ability was the result of multiple factors of cognitive abilities rather than a single overarching factor.

Despite Thurstone's assertion that intelligence was comprised of seven independent abilities, subsequent studies failed to replicate his findings. Rather, later studies showed that the original factors that Thurston obtained were less orthogonal than originally believed (Thurstone & Thurstone, 1941). Thus, after noting the intercorrelations obtained between mental ability measures specifically designed to assess discrete facets of cognitive ability, Thurston (1947) doubted the possibility that an orthogonal structure of intelligence could be developed that did not capture *g*, reconciling his ideas with those of Spearman.

Structure-of-Intellect. Despite Thurstone's assertion, other theorists have denied the existence of *g*. For instance, based on his own factor analytic research, Guilford (1967, 1988) developed a model that eliminated the role of *g* in explaining performance on intelligence measures. Guilford's Structure-of-Intellect (S-I) is a box-like schema from which intellectual traits are classified along three dimensions: Operations, Contents, and Products. Operations represent the mental activities or processes that are performed by the individual. Operations can be further classified as cognition, memory recording, memory retention, divergent production, convergent production, and evaluation. Contents represent the information or materials that receive the operations. Contents include auditory, visual, semantic, symbolic, and behavioral information. Products represent the various forms in which content may be processed. Products can be further classified as units, classes, relations, systems, transformations, and implications.

Each factor of cognitive ability in Guilford's S-I model is derived from the sub-classifications of the Operations, Contents, and Products dimensions. Since at least one factor is expected from each cell in the schema, 180 (6 x 5 x 6 = 180) or more factors constitute intellectual ability. As such, abilities represent a specific operation, in a specific content area, leading to a specific output (e.g., Evaluation of Semantic Implications). Since Guilford considered the factors that were produced from the S-I model to be orthogonal, he rejected the value of *g* and hierarchical relationships of mental abilities. Like Thurstone's model, the S-I model was derived from an orthogonal rotation of test scores (Guilford, 1967). However, unlike other factor analytic theories, the S-I theory of intelligence was derived from a theoretical basis and tests were then constructed to measure the hypothesized components.

However, the S-I model failed to gain an influential foothold in cognitive ability testing (Carroll, 1993). Likewise, re-analyses of Guilford's factor-analytic data indicate that other models provide better fit to the data, including randomly generated models. As such, Carroll (1993) described the considerable amount of attention paid to the S-I model as *disturbing* and as providing the impression that the model is a widely accepted and valid theory of cognitive ability, which it is not.

Gf-Gc. In contrast to the S-I theory, a model that is widely accepted is the *Gf-Gc* theory forwarded by Cattell (1941). Based on the works of Thurstone in the 1930s, the original *Gf-Gc* theory suggests that intellectual ability is comprised of two primary abilities: Fluid and Crystalized Intelligences.

Fluid Intelligence (*Gf*) consists of the focused attention to process information and solve problems that cannot be performed automatically and/or are independent of any learned information (Schneider & McGrew, 2012). Since the abilities that are associated with *Gf* are biologically rooted, they are thought to be culture-free, non-verbal, and independent of any form of instruction allowing individuals to adapt to new situations and learn from their environments (Cattell, 1957, 1971). Thus, individuals who possess high levels of *Gf* are able to act quickly and encode short-term memories that enable abstract problem solving. The mental operations associated with *Gf* that promote problem solving include recognizing and transforming information and concepts, perceiving relationships among patterns, drawing inferences or otherwise extrapolating answers, and comprehending the implications of the solution reached. Inductive and deductive reasoning are the defining characteristics of fluid intelligence best measured through tasks including figural matrices, number series, analogical reasoning, and figural analyses (Sattler, 2001).

Crystalized Intelligence (*Gc*) consists of the acquired skills and knowledge that are derived from one's experience and valued by one's culture (Schneider & McGrew, 2012). As described by Horn and Blankson (2005), the abilities associated with *Gc* are verbally based, developed through an investment of mental energies into educational and other life experiences. The types of knowledge that are subsumed by *Gc* include both static declarative (e.g., factual information, comprehension, concepts, rules, and relationships) and dynamic procedural (e.g., process of reasoning based on previously learned information). As such, *Gc* is not only a repository of information, but is also a set of processing abilities wherein memory retrieval and the application of general knowledge are components.

Cattell-Horn. Through his own factor analytic research, Horn (1968, 1988, 1991) expanded on Cattell's dichotomous *Gf-Gc* model, adding several additional factors: visual perception or processing (*Gv*), speed of processing (*Gs*), short-term memory (*Gsm*), long-term memory (*Glr*), auditory processing ability (*Ga*). Later, Horn added factors representing reaction time and decision speed (*Gt*), quantitative (*Gq*), and broad reading-writing (*Grw*) abilities. This conglomerate eight-factor model became known as the Cattell-Horn theory (Horn, 1991).

Three-Stratum Theory. Carroll (1993) reported the exploratory factor analytic results of over 460 datasets, building on the research of influential theorists such as Cattell, Horn, Thurstone, and Thorndike. The magnitude and comprehensiveness of this analysis was not lost on other researchers. As indicated by Jensen (2004), "Carroll's magnum opus thus distills and synthesizes the results of a century of factor analyses of mental tests. It is virtually the grand finale of the era of psychometric description and taxonomy of human cognitive abilities. It is unlikely that his monumental feat will ever be attempted again by anyone, or that it could be much improved on" (p. 5).

Carroll's influential Three-Stratum theory of intelligence differentiates factors and abilities into three hierarchically arranged levels (Carroll, 1997). In geological terms, a stratum is a bed of sedimentary rock or soil that distinguishes itself from adjacent strata. Similarly, Carroll proposed that intelligence is best modeled in hierarchical terms. The top stratum, Stratum III, is *g* or general intellectual ability. As represented by Spearman (1904), *g* is a broad processing ability that is behind all higher-order thinking and subsumes the other two strata in the models. Known as broad or Stratum II abilities, the second stratum is comprised of eight abilities, incorporating Cattell's fluid (*Gf*) and crystalized (*Gc*) intelligences, along with broad visual perception (G*v*), broad auditory perception (*Ga*), broad retrieval capacity (*Gr*), broad cognitive speediness (*Gs*), processing/decision speed (*Gt*), and general memory and learning (*Gy*), each requiring differing processes, tasks, and content. These abilities are the most recognized and prominent abilities in Carroll's model, representing "basic constitutional and long standing characteristics of individuals that can govern or influence a wide variety of behaviors in a given domain" (Carroll, 1993; p. 634). Below each Stratum II ability lays the 69 Stratum I *level factors* or *speed factors* that are associated with a specific Stratum II ability (Jensen, 1998). These narrow abilities "...represent greater specializations of abilities, often in quite specific ways that reflect the effects of experience and learning, or the adoption of particular strategies of performance" (Carroll, 1993, p. 634). Although other theorist such as Burt (1949) and Vernon (1950) proposed hierarchical models of intellectual ability, Carroll's model was the first "empirically based taxonomy of cognitive ability... presented in a single organized framework" (McGrew, 2009, p. 2).

Importantly, the abilities in Carroll's model exhibit positive relationships with one another. As such, the mutual relationships shared between the narrow Stratum I abilities gives rise to the broader Stratum II abilities. Likewise, the positive relationships that associate Stratum II abilities allows for the approximation of the *g*-factor at Stratum III. Although these positive relationships indicate that the abilities are not completely orthogonal to one another, a vast amount of research indicates that the factors can be consistently differentiated from one another, thus indicating that they are in fact unique facets of cognitive ability (Keith & Reynolds, 2010).

Cattell-Horn-Carroll. Carroll (1993) stated that the Cattell-Horn *Gf-Gc* model "appears to offer the most well-founded and reasonable approach to an acceptable theory of the structure of cognitive abilities" (p. 62). Perhaps due to his admiration of the work of Cattell and Horn, the Carroll's Three Stratum and the Cattell-Horn models are quite similar. For example, both of the proposed models contain broad abilities that subsume narrower abilities. Likewise, both models share similar classifications of these abilities. However, the models are distinguished from one another. Several of the differences between the models involve the definitions attributed to specific abilities and the groupings of narrower facets. The biggest difference between the models is existence of *g*. That is, Carroll's model suggests that an overarching *g*-factor subsumes narrower abilities, while the Cattell-Horn model does not include a *g*-factor.

Despite the differences embodied by the Cattell-Horn and Carroll models, researchers recognized the need for a common framework to describe, organize, select, and interpret assessments and assessment batteries. To meet this need, McGrew (1997) proposed a hybrid model combining the Cattell-Horn and Carroll models in to what became known as the Cattell-Horn-Carroll (CHC) theory, with the order of the names reflecting the chronological order in which the theorists made their contributions. As such, CHC theory represents over 60 years of factor analytic research of cognitive ability.

The CHC model is arranged in three hierarchal levels. Like Carroll's Three-Stratum theory, at Stratum III, the top level, the general factor of intelligence or *g* resides. Stratum II contains the broad cognitive abilities while the narrow abilities lie at the bottom level in Stratum I. In its original configuration, CHC theory contained 10 broad cognitive abilities and over 70 narrow abilities. However, CHC theory is not static. Rather, CHC is continuously refined, reorganized, and restructured as additional research is conducted (McGrew & Flanagan, 1998; Flanagan, 2000). As stated by Jensen (2004), CHC is "an open-ended empirical theory to which future tests of as yet unmeasured or unknown abilities could possibly result in additional factors at one or more levels in Carroll's hierarchy" (p. 5). Carroll (2005) reiterates this point noting that CHC most assuredly contains errors that may be rectified through continued research. In its current form, CHC theory consists of 16 broad stratum abilities and over 80 narrow abilities (Schneider & McGrew, 2012). The 16 broad stratum abilities of CHC currently include

Fluid Intelligence (*Gf*), Crystalized Intelligence (*Gc*), General (Domain-Specific) Knowledge (*Gkn*), Quantitative Knowledge (*Gq*), Reading/Writing Ability (*Grw*), Short-Term Memory (*Gsm*), Long-Term Storage and Retrieval (*Glr*), Visual Processing (*Gv*), Auditory Processing (*Ga*), Olfactory Abilities (*Go*), Tactile Abilities (*Gh*), Psychomotor Abilities (*Gp*), Kinesthetic Abilities (*Gk*), Processing Speed (*Gs*), Decision Speed/Reaction Time (*Gt*), and Psychomotor Speed (*Gps*). However, of the Stratum II abilities, *Gf* and *Gc* are the most related to *g* (Carroll, 2003).

Although the CHC model was forwarded by McGrew (1997) to pragmatically classify narrow cognitive ability measures that are contained in individually administered intellectual assessments, this model is the most theoretically sound and empirically supported model of intelligence available (Ackerman & Heggestad, 1997; McGrew, 2009; Stankov, 2000). For instance, the factor structure of CHC is supported by factor analytic evidence that not only demonstrates the consistency of the factors derived, but the invariance of the three-stratum factor structure across one's life (Bickley, Keith, & Wolfle, 1995) and across gender, ethnic, and cultural groups (Carroll, 1993). Likewise, evidence provided from studies of developmental, neurocognitive, and heritability lend support to the CHC model (Horn & Blankson, 2005). As such, the CHC model is backed by a more extensive array of validation evidence than any other modern theory of cognitive ability (Schneider & McGrew, 2012).

The core practice in scientific fields is the classification of empirical observations (Bailey, 1994). As argued by Miller (1996), useful taxonomies draw distinctions of conceptual importance, raise contrasts that enable empirical advancement, and possess elements that form a coherent whole. Given the substantial amount of evidence supporting the structure of the CHC model, the value of such a model is the common framework that allows practitioners to think alike regarding the measurement of cognitive abilities and the usefulness of the broad vs. narrow facets. The CHC model is particularly relevant to the area of school psychology and psychoeducational assessment as several measures of cognitive abilities have incorporated CHC as a theoretical foundation. For instance, CHC provides researchers a means to design and evaluate cognitive assessments and a common language for describing research findings that stimulates the empirical investigation of the structure and nature of cognitive abilities (Keith & Reynolds, 2010).

Theoretical Approaches to Intelligence

Despite the advances that have been made through the factor analysis of test scores, some researchers maintain that theories of intelligence that are derived from such exploratory analyses fail to capture the extent of cognitive functioning. In contrast to factor analytical accounts, a variety of theoretical frameworks have been constructed that purport to better conceptualize and measure intellectual ability.

Successful Intelligence. While Boring (1923) famously stated that intelligence is what the tests test, Sternberg was dissuaded by conventional measures of intelligence that consisted solely of measures of analytical and memory items. Rather, Sternberg (2005) proposed a competing information-processing model that has received considerable attention. The Successful Intelligence theory, also known as Triarchic theory of intelligence (Sternberg, Grigorenko, Ferrari, & Clinkenbeard, 1999), suggests that traditional measures of cognitive ability focus too much on analytical abilities, ignoring creativity and practical thinking that allow an individual to deal effectively with the world. Sternberg acknowledged that the measures of analytical abilities and memory used in traditional measures of cognitive ability are important indicators of success in academic environments. However, to achieve success "one needs not only to remember and analyze concepts; also one needs to be able to generate and apply them" (Sternberg, 2005; p. 190). As Sternberg argued, there is a multitude of ways for someone to be successful at their job. That is, people achieve goals though selecting, shaping, and adapting to their environment and contexts. What works for one person may not work for another, but successful people modify their environments or circumstances to exploit their skills and mitigate or eliminate their weaknesses. In contrast, the unsuccessful fail to capitalize on their limited talents.

Since each path to success is different, Sternberg (2005) argued that what is meant by *intelligence* will have a different meaning to each individual. Rather than an overarching general intelligence or *g*, success is achieved through the combination and utilization of varied forms of thinking, namely Analytical, Creative, and Practical intelligences. Analytical intelligence is used to analyze, judge, evaluate, compare, and contrast relatively familiar, but abstract problems. Creative intelligence is used to cope with relative novelty. Practical intelligence is used to select, shape, and adapt environments to suit oneself. As suggested by Sternberg (2005), the strong relationship noted between measures of *g* and academic success is in part due to the failure of traditional measures to assess creative and practical intelligence.

Core to Sternberg's theory are the universal component processes that contribute to the information processing required for analytical, creative, and practical thinking. A component is defined as "an elementary information process that operates upon internal representations of objects or symbols" (Sternberg, 1977; p. 65). Components are thought to be universally applied across cultures, but their application will depend on the nature of the problem faced (Sternberg, 2004). The theory of successful intelligence proposes that three component processes underlie human intellect: Metacomponents, Performance components, and Knowledge-acquisition components. Metacomponents are the executive processes that are responsible for the identification of problems, strategizing a solution, monitoring progress towards a goal, and evaluating the effectiveness of the resultant solution. Performance components set the plans of the metacomponents into action. Knowledge-acquisition components are used to learn new declarative information and/or how to solve problems.

Although Sternberg (2005) has supplied evidence to support the efficacy of Successful Intelligence theory, the procedures used in these studies have met with strong criticisms, limiting the interpretability and veracity of evidence provided. For example, the measures of the Sternberg Triarchic Abilities Test (STAT; Sternberg et al., 1999) are inherently unreliable (Brody, 2003). When corrected for range and unreliability, the correlations between the STAT and measures of *g* are highly related. Thus, Successful Intelligence appears to substantially related to *g*. Brody further demonstrates that Sternberg's three forms of intelligence correlate at .62 or higher, indicating substantial overlap between the supposedly independent factors. Likewise, the measures fail to demonstrate convergent validity with other measures with which they should be theoretically related. Gottfredson (2002) strongly criticized Sternberg's assertions, indicating that the authors "can support their...major theoretical propositions only by ignoring the most relevant evidence on *g* and making implausible claims about practical intelligence" (p. 3). As such, the conclusions drawn by Sternberg that Successful Intelligence theory is measuring attributes independent of *g* are highly suspect.

Multiple Intelligences. However, Stenberg is not alone in the assertion that intelligence is not well explained by *g*. Dismayed by the fact that traditional measures of intelligence and academics in general focus predominantly on logical and linguistic abilities, Gardner (2011) proposed his theory of Multiple Intelligences (TMI). According to Gardner's TMI, intelligence is often defined too narrowly, including only those capacities that are important for academic success. Rather, Gardner suggests that intelligence is better represented by a spectrum of abilities. In practice, people draw on one or more of these abilities at a time to produce outcomes or "end states." For instance, Gardner and Hatch (1989) argue that few occupations rely on a single form of intelligence. A surgeon, for example, must be able to not only solve problems as they arise in an operating room, but also possess the manual dexterity to manipulate a scalpel to correct the issue. As such, the surgeon is drawing on multiple forms of intelligence to influence success on-the-job.

In its current form, TMI consists of nine distinct but closely related intelligences: Verbal-Linguistic, Logical-Mathematical, Spatial-Visual, Body-Kinesthetic, Musical, Naturalistic, Existential, Interpersonal, and Intrapersonal (Gardner, 2011). Since each form of intelligence is hypothesized to be independent of all of the others, a person can be described by a unique intellectual profile of the nine intelligences, highlighting one's intellectual strengths and weaknesses. Gardner advocates that academic environments and curricula be tailored to suit the needs of individual students and their pattern of intellectual abilities. Thus, schools will be equipped to identify and remediate a child's weaker intelligence(s). However, since all intelligences are also interrelated, as one becomes more proficient in a specific area, all areas of intelligence are enhanced. As such, Gardner believes that schools should be filled with a variety of interesting toys, books, games, and objects which can be manipulated and explored, thus providing students with a multitude of options to explore the world and enhance their intellectual capacity. Many schools have adopted the principles of TMI, producing a substantial impact on the American educational system (Lubinski & Benbow, 1995).

Despite the acclaim and attention that Gardner's TMI theory has obtained in academic settings, critics remain unconvinced of the merits of TMI. From a theoretical perspective, the fact that all of the forms of intelligence are supposedly interrelated and performance in one area can promote growth in another supports the influence of an overall *g* factor. In fact, Gardner's intelligences correlate well with standard measures of intelligence (e.g., the Wonderlic Personnel Test) and form a substantial *g*-factor (Visser, Ashton, & Vernon, 2006). From a psychometric perspective, no empirical evidence has been provided to demonstrate the claims made by Gardner, nor has the theory been specified in enough detail to be effectively evaluated (Hunt, 2001). In Lubinski & Benbow's (1995) critical review of TMI, the authors note that Gardner has gone to great lengths to describe the various forms of intelligence reinforcing the face validity of the theory. However, Gardner has failed to demonstrate that these intelligences are related to real world outcomes. Likewise, Gardner has failed to provide reliability estimates for any of his scales. Thus, any inferences that can be drawn from his measures' relationship to outcomes are suspect at best. As such, until meaningful evidence supporting the reliability and validity of TMI is available, it poses limited utility.

Planning, Attention, Simultaneous, and Successive. Derived from Luria's (1966) organization of brain functioning, the Planning, Attention, Simultaneous, and Successive (PASS) cognitive processing theory of intelligence (Naglieri & Das, 1997; Naglieri, Das, & Goldstein, 2014) focuses on how information is processed rather than the kinds of information that are processed. The PASS theory represents the integration of cognitive and neuropsychological research, positing four interrelated, yet distinct neurocognitive abilities associated with various regions of the brain. Planning is the ability to control and direct one's thoughts and actions to obtain an efficient solution to a problem. Attention is the ability to direct one's mental energy toward a target stimulus while inhibiting responses to competing stimuli. Simultaneous processing is the ability to integrate disparate parts into groups or an integrated whole. Successive processing is the ability to recognize and sequential or serially order information. Since various parts of the brain are involved in different kinds of information processing, the PASS theory does not allow for a higher order *g*-factor.

Proponents of the PASS model argue that planning has not been adequately measured by other intellectual instruments, resulting in the misspecification of specific academic deficits associated with specific cognitive problems. As such, the Cognitive Assessment System (CAS; Naglieri & Das, 1997) and revised version (CAS-2; Naglieri et al., 2014) were developed, explicitly measuring processes that other psychometrically derived measures of intelligence have failed to assess.

However, despite the validation evidence supporting the criterion and construct-related validity of the CAS and CAS-2, evidence indicates that the abilities measured by the PASS model are more consistent with the CHC model (Keith &

Reynolds, 2010; Kranzler & Keith, 1999). Namely, when subjected to confirmatory factor analysis, the PASS model produces a poor fit to the data. When competing theoretical models are applied to the same data, the best model fit is provided by a third-order hierarchical model with a general factor (*g*) of intelligence at the top, an intermediate Planning/Attention factor, and four narrow facets associated with the PASS abilities. Moreover, although the PASS model was born out of strong theoretical origins and designed to measure non-*g* related abilities, when students were administered measures, the *g*-factor derived from the CAS and *g*-factor of the Woodcock-Johnson-III correlated at .98 suggesting that the two are nearly indistinguishable (Keith, Kranzler, & Flanagan, 2001). As such, the CAS appears to have the same measurement characteristics of the widely replicated CHC models.

The Nature of *g* and *Gf*

Despite the attention that the theoretical accounts of intelligence of have received, thus far none have shown the utility exhibited by the CHC model. Moreover, although theoretical models of intelligence deny the existence of an overarching general mental ability, when critically analyzed, these models show substantial relationships with *g* (Brody, 2003; Keith et al., 2001; Visser et al., 2006). As such, despite attempts to measure aspects of intelligence that are independent of *g*, the construct continues to emerge. Thus, the measurement of *g* remains the best estimate of GMA.

As noted earlier, the strength of the CHC model rests on its ability to serve as a bridge from theory to practice, guiding the design and selection of cognitive ability instruments and batteries capturing the qualities that relate to our current understanding of intellectual functioning. Through multiple replications and substantial validation efforts, the CHC model has emerged as the most complete, structurally sound, and valid model of cognitive ability. Thus, most new measures of cognitive abilities are based on the CHC model, acknowledging its fidelity (Keith & Reynolds, 2010).

In order to capture the complete range of mental abilities, the CHC prescribes a battery of assessments that provide the mosaic measurement of intellect. Given that the CHC model currently denotes 16 broad abilities and a host of narrower facets, a complete CHC-based assessment battery would necessitate a lengthy administration. In developing measures of cognitive ability or any other psychological construct, test developers are faced with a tradeoff between thoroughness and accuracy. For example, the length of a test is directly related to its reliability (DeVellis, 2012). Using classical test theory, longer tests are inherently more reliable since they capture items of increasing redundancy. In doing so, error variance is reduced, providing a more focused and hence, reliable measure. Thus, it would seem that an infinitely long measure would be desirable. Yet, due to test taker fatigue, it is a commonly recommended practice in psychometrics to consider reducing the number of items in a scale once the reliability coefficient reaches an adequate threshold. This same sentiment is a consideration for assessment batteries where a universe of items is possible, but the administration of a large number of items will result in test taker fatigue. A developer of a test of cognitive ability must balance the creation of a measure that covers the breadth of mental abilities with the expediency of producing useful and valid results.

The same tradeoff must be made when developing a more focused measure of general cognitive ability. Spearman's psychometric *g* is implied through the positive correlations among mental ability measures. Therefore, it is not possible to measure *g* with any single measure. Rather, intelligence is approximated through the aggregation of highly *g*-saturated measures (Carroll, 1993). Similarly, Ackerman et al. (2005) notes that the uses of a single measure of cognitive ability raises the possibility that *s* factors will be captured in addition to *g*. Despite this, Ree, Earles, and Teachout (1994) have questioned the practicality and necessity of obtaining a comprehensive estimate of cognitive ability through an extensive battery of mental measures. Likewise, Spearman recommended that the use of a single highly *g*-saturated measure is preferable to the use of several heterogeneous cognitive measures that capture a variety of abilities (Spearman, 1927).

At Stratum II of the CHC model, *Gf* and *Gc* are the two most highly *g*-saturated broad abilities, with *Gf* more closely relate to *g* (Carroll, 1993). As originally conceived by Cattell (1971), *Gf* is used as a means to enhance other mental abilities, such as the accumulation of *Gc* through focused attention. As asserted by Gustafsson (1984, 1989, 2001) and others, *Gf* is indistinguishable from *g* when subjected to confirmatory factor analyses. This suggests that fluid abilities represent the foundation of general intelligence. For instance, Arendasy, Hergovich, and Sommer (2008) tested the *g*-saturation of the Stratum II factors, finding that *Gf* is virtually identical to psychometric *g*. As such, *Gf* measures produce large *g*-saturations even without averaging over several subtests. Therefore, *Gf* can be thought of as the raw horsepower of cognitive functioning, indicative of general mental ability.

However, contrary to the contention made by Gustafsson (1984, 1989, 2001), *g* and *Gf* do not appear to be the same construct. Rather, Carroll's (2003) analysis demonstrates that *Gf* "...is significantly separate and different from *g*, tending to disconfirm any view that *Gf* is identical to *g*" (p. 14). However, Carroll points out that the issue has not been completely resolved, speculating that it is likely difficult to develop a measure of *Gf* that is reliably independent of *g*, stating that "better tests of *Gf* are needed to establish this factor as linearly independent of factor *g*, if indeed this is possible..." (Carroll, 2003, p. 19). Whatever qualities are associated with *g*, measures of non-verbal reasoning and novel problem solving through the use of spatial elements and inductive reasoning seem to best capture it. As such, tests of *Gf* are thought of as good approximations of Spearman's *g* (Ackerman et al., 2005).

While the types of items that best capture *Gf* are known, substantial confusion surrounds the measurement of the *Gc* construct. Adding to the confusion related to its measurement, various terms such as crystalized intelligence, comprehension, and academic achievement are used by professionals to describe the construct (Keith & Reynolds, 2010). *Gc* implies a depth of knowledge that would describe someone who possesses a vast repository of information. However, as pointed out by Horn and McArdle (2007), measures of *Gc* rarely measure beyond surface knowledge. Likewise, individuals who score well on *Gc* measures tend to have a wide breadth of knowledge. Therefore, it is impossible to distinguish the *Gc* abilities of experts and those of individuals who have a superficial knowledge on a wide variety of topics based solely on *Gc* scores. As such, *Gc* scores may not be as useful as believed.

There is also reason to believe that *Gc* is becoming less relevant as technological advances permeate society. Although *Gc* measures are good indicators of academic and business success from which hard work can positively influence test scores, advances in computer technology can store far more information than any one person can accumulate, holding it accurately, securely, and cheaply. That is, while crystalized knowledge is a repository of information that comes with experience and education, Furnham (2008) argues that the future belongs to quick-witted individuals who are able to think on their feet, adapt to rapidly changing circumstances, and reason effectively.

Taken together, the evidence suggests that at its core, *g* is best tapped by *Gf* measures. As indicated by Cattell (1971), *Gf* is the governor of intellectual ability. Lesser abilities are dependent on the investment of *Gf*. Likewise, due to its high *g*-saturation, tests consisting of reasoning and novel problem solving abilities that are associated with *Gf* should be the predominant item types for brief measures of GMA (Arendasy & Sommer, 2012; Carroll, 1993, 2003; Gustafsson, 1984, 1989, 2001).

The Power of Intelligence

The relationship between intellect and success is most noticeable in academic settings (Ones, Viswesvaran, & Dilchert, 2006) where measures of intelligence are used in predicting exam scores, amount of learning, and academic success in schools and universities, regardless of the subject or specialty. Likewise, education has a strong reciprocal impact on intelligence (Ceci, 1991). This relationship is important because it leads to compounding life-advantages. As noted by Feldman (1966), individuals with more education seek out and acquire more information. For instance, individuals who obtain higher levels of education use periodicals such as books, newspapers, and magazines to a greater extent than their less educated peers. This is the precise reason that higher cognitive ability promotes additional learning; higher levels of *g* are associated with increased exposure to information, which is in turn exploited to a greater degree (Gottfredson, 2004).

Skeptics of *g* argue that intelligence is little more than an academic skill (Sternberg et al., 2000). However, the non-academic value of *g* has a clear and predicable influence on occupational attainment, social life, and even one's life span (Deary, 2004; Gottfredson, 1997; Lubinski, 2004; O'Toole & Stankov, 1992; Schmidt & Hunter, 2004). This general cognitive ability or ability to deal effectively with cognitive complexity (Gottfredson, 1998) is the hallmark of intelligence across contexts, allowing for the processing of information of any sort, constituting the backbone of human mental ability (Gottfredson, 2004). For example, the effects that cognitive ability has on problem solving and learning in everyday situations are robust (Gottfredson, 2002). Intellect is shown to predict important life outcomes such as incarceration, poverty, health, and mortality due to engaging in risky health behaviors (Gottfredson, 2004). Likewise, a variety of important occupation-related outcomes such as job performance, income level, and occupational attainment are predicted by intelligence (Schmidt & Hunter, 2004). Thus, the effects of *g* are pervasive because it is highly transportable. In other words, there is a linear relationship between an individual's level of *g* and performance in school, work, and social situations (Gottfredson, 2004). The general effect of this relationship results in greater life success, producing dividends across situations, time, and cultures (Gottfredson, 2004; Nettlebeck & Wilson, 2005).

As argued by Gottfredson (2004), life itself can be thought of as a cognitive abilities test. There are virtually no aspects of our daily lives, no matter how trivial, that do not require the ability to reason, plan, or solve problems. For example, everyday activities such as reading the directions listed in a recipe, determining how much to tip a waiter at lunch, or reading a map exert a cognitive load, requiring the ability to reason and solve problems. As such, there are few situations where being less intelligent is advantageous. However, the choices that are made on a daily basis produce compounding returns that can result in large differences (Gottfredson, 2004). For example, there are many reasons why someone may fail a given task that are unrelated to cognitive ability, such as misreading the headline of a newspaper or becoming lost while looking at a street map in an unfamiliar city. The person may be distracted, tired, and/or hungry, but these effects tend to be transient and unreliable. Conversely, the effects of cognitive ability are pervasive and fairly consistent across life situations. Just as casinos know that small gaming odds in their favor can produce huge dividends over time, small edges in cognitive ability aggregate and produce large effects over a lifetime (Gordon, Lewis, & Quigley, 1988). Individuals with higher levels of cognitive ability make better judgments by exercising better problem-solving and reasoning abilities in everyday situations (e.g., managing finances or reading a map). In contrast, individuals who are less adept at planning and budgeting slowly slip behind others who initially began with the same resources. As such, when this slippage occurs, it occurs in many realms of life, producing pronounced effects. Moreover, these slippage effects are expected to become more pronounced as the world becomes increasingly connected (Cascio & Aguinis, 2005). Technological and societal advances have amplified the complexity of daily life, increasing the number of choices that must be made, placing a premium on cognitive ability (Gottfredson, 2003). As such, intelligence is more pervasive and inclusive than a narrow abstract skill that allows one student to shine academically where another languishes. Rather, cognitive ability is a broad intellectual capacity to interact with the world effectively.

Intelligence in the Workplace

Measures of cognitive ability are among the most predictively valid employee selection measures available to organizations (Schmidt & Hunter, 1998). Based on data collected on over 32,000 employees in a variety of jobs conducted for the U.S. Department of Labor (Hunter, 1980; Hunter & Hunter, 1984), meta-analytic evidence indicates that the overall predictive validity of cognitive ability to job performance is .51 for jobs of median complexity (Schmidt & Hunter, 1998). Since performance data are only available for those applicants who are hired, Hunter, Schmidt, & Le (2006) estimate that the true validity coefficient of cognitive ability may be well over .60 once corrected for range restrictions.

The predictive validity of cognitive ability also rises as job complexity increases (Ones et al., 2006; Schmidt, Hunter, & Pearlman, 1981). The most complex jobs are those that are abstract, cannot be routinized, and are autonomous, thus allowing workers to exercise more discretion (Schmidt & Hunter, 2004). As such, complex jobs place a premium on workers' ability to reason, solve problems, and make judgments without supervision. For instance, Hunter and Hunter (1984) reported that the highest mean validity of cognitive ability that they found was for professional-managerial jobs (.58), followed by highly technical jobs (.56), medium complexity jobs (.51), semi-skilled jobs (.40), and unskilled laborer (.23). Conversely, only in the lowest, least complex, and most routinized positions, do constructs such as tenure and psychomotor abilities better predict on-the-job performance than cognitive ability (Gottfredson, 2002). Moreover, since the utility of a selection device is directly tied to its validity (Schmidt & Hunter, 1998), the dividends of using valid selection devices can reach millions of dollars over time, whereas organizations that make poor decisions using invalid instruments stand to lose millions of dollars in reduced production.

The relationship between possessing a high intellect and employment success has been known to researchers for quite some time. For example, Harrell and Harrell (1945) noted that employees of lower intelligence were less likely to rise up the organizational hierarchy to obtain the prestigious "white-collar" positions. This assertion is buttressed by U.S. Employment Service data, showing a strong correlation (.72) between cognitive ability and job level (Jensen, 1998). However, cognitive ability has also been shown to predict job movement into positions of either higher or lower complexity. For instance, using a sample of 3,887 young adults, Wilk, Desmarais, and Sackett (1995) demonstrated that cognitive ability measured in 1980 predicted job movement over a five-year period (1982-1987). Specifically, the results of this study show that individuals with greater cognitive ability tended to move up the organizational hierarchy while those with less cognitive ability moved down. In a follow-up study, Wilk and Sackett (1996) found that job mobility was predicted by the congruence of cognitive ability and job complexity. Individuals who possessed a cognitive ability that was greater than the complexity of their job tended to move into positions of greater complexity. Conversely, individuals who possessed a cognitive ability that was less than the level of complexity of the job tended to move into less complex positions. Likewise, greater variability in cognitive ability scores are seen in less complex positions, but a consistent upper range of scores are found across occupations suggesting a minimal level of intellect is required as one rises in the hierarchy (Harrell & Harrell, 1945). Thus, while people of high intelligence occupy low complexity jobs, access to higher-level positions require greater levels of cognitive ability. This point is also echoed by Gottfredson (2004) who indicates that in the United States, the most coveted and highest paying jobs go to the cognitive elite, while the less cognitively endowed workers are doomed to a life of menial labor and low pay in our informationally based economy. As such, it quite literally pays to be smart.

The relationship between cognitive ability and job success is not limited to the United States. The findings of Schmidt and Hunter (1998) and Ones et al. (2006) are reinforced by Salgado, Anderson, Moscoso, Bertua, and De Fruyt (2003) who conducted a similar meta-analytic investigation of the relationship between intelligence and on-the-job performance in a European sample consisting of over 25,000 workers. After corrections were made for measurement error, the findings of this study suggest that the operational validity of cognitive ability is .62, but the value was smaller for specific forms of intelligence. Similarly, Bertua, Anderson, and Salgado (2005) and Hulsheger, Maier, and Stumpp (2007) examined cognitive ability in British and German samples, respectively. Consistent with previous findings, the data revealed that as job complexity increases, the predictive validity of cognitive ability increases. In sum, not only do the data suggest that cognitive ability is the single best predictor of occupational success for any occupation or industry, but cognitive ability is the best predictor of job performance internationally too.

Why Does Cognitive Ability Affect Performance?

Although the link between cognitive ability and job performance is strong, why is the relationship so robust? As suggested by Schmidt and Hunter (2004), cognitive ability is thought to influence performance indirectly. Cognitive ability allows for the faster and more thorough absorption of essential job knowledge. In turn, the information learned is exploited to a greater degree, allowing employees to go beyond their current knowledge of the job and to make judgments in novel and changing situations. Similarly, Borman, Hanson, Oppler, Pulakis, and White (1993) suggest that the relationship between intelligence and job performance is mediated. Higher intelligence results in individuals having more opportunities to obtain additional job experience. The experience gained then leads to additional job knowledge.

Given the vast amount of evidence showing the relationship between cognitive ability and job performance, it is no surprise that higher cognitive ability is also related to employee training outcomes. Schmidt & Hunter (1998) report that no other measure has the predictive power of cognitive ability (r = .56) in predicting training success. Moreover, similar results (r = .54) were found when intelligence was used to predict training performance in European samples (Salgado et al., 2003). Thus, when an employer uses cognitive ability as a selection measure, the employer is also selecting individuals who are better able to rapidly learn on the job. Consequently, Schmidt and Hunter (1998) recommend that cognitive ability should be considered the primary tool for selection decisions.

Adverse Impact

Despite the high predictive validity and other advantages associated with cognitive ability testing, organizations remain hesitant to use such devices to make employment decisions due to the consistent and near universal finding that cognitive ability measures produce differential scoring across racial subgroups (Campbell, 1996; Hartigan & Wigdor, 1989; Sackett & Ellingson, 1997). Specifically, lower than average scores are observed in African-American and Hispanic samples, while groups such as

Asian-Americans tend score higher than average (Rushton & Jensen, 2005). Comparing racial subgroups, African-Americans score about 1 standard deviation in the population lower than Caucasians, while Japanese and Chinese samples obtain the highest scores.

As a result, occupational outcomes can be partially explained by the gap in cognitive ability scores. According to Gottfredson (2002), only about 22% of Caucasians and 59% of African-Americans produce cognitive ability scores below 90. As such, fewer African-Americans are considered competitive for mid-level jobs and trades such as firefighters and clerical workers. The average cognitive ability score for incumbent to these types of jobs is one standard deviation above the average score of African-Americans. Conversely, on the other end of the continuum, the ratio of African-Americans to Caucasians producing cognitive ability scores of 125 or greater is 1:30 the average for the most socially desirable professional positions such as lawyers, physicians, and engineers.

Despite the racial gap in cognitive ability scores noted in the general population, McDaniel and Banks (2010) argue that for two reasons, these differences should be less pronounced when actual job applicants compete for jobs. First, individuals at the lowest levels of cognitive ability do not have the mental capabilities to perform effectively in common jobs and therefore are not job applicants. Second, job applicants must meet minimal job requirements (e.g., education and experience) when applying for a position. Pools of qualified job applicants who have obtained formal education and possess the requisite relevant job experience are more likely to be homogenous in respect to cognitive ability than random samples drawn from the general population. Therefore, due to the pre-screening of applicant qualifications, larger racial gaps in cognitive ability scores are more likely to be found in lower level jobs that do not require lofty education and experience requirements. Conversely, smaller racial gaps should be noted in positions that require high levels of education and experience. Substantiation of this assertion is provided by Roth, BeVier, Bobko, Switzer, and Tyler (2001) who note that the standardized mean difference between white and black applicants shrinks from 1.0 in the general population, to .86 for low complexity jobs, .72 for medium-complexity jobs, and .63 for high-complexity jobs.

Despite the reduction in test score differences seen across job complexity, the observed deviations can still cause disparate hiring practices if cognitive ability measures were used as the sole selection instrument. As such, challenges to the legality of cognitive ability testing began in 1971 with the influential Griggs v. Duke Power (1971) case. As a result of this case, the Supreme Court ruled that when a selection procedure or device produces adverse impact against a protected group, the organization must be able to demonstrate that the use of the measure is a "business necessity," imperative to organizational functioning and operation. However, as indicated by Grover (1996), courts have generally held rather narrow interpretations on what constitutes a business necessity that hamper the use of alternative selection options. As a result, many organizations curtailed their usage of cognitive ability measures in making employment decisions.

The Measurement of Cognitive Ability

Despite the legal challenges that surround cognitive-ability measures, Nisbett et al. (2012) believe that the measurement of intelligence is one of the greatest accomplishments of psychology. As stated by DeVellis (2012), "measurement is a fundamental activity of science" (p. 2). As such, despite cognitive theorists who have devised theoretical accounts of intelligence that elude measurement, Cronbach (1990) notes, “If a thing exists, it exists in some amount. If it exists in some amount, it can be measured” (p. 34).

The History of Testing and Measurement

The historical roots of measurement stretch into antiquity. Duncan (1984) argues that measurement is an inherently social process that emerged in ancient humans as a means to overcome the problems faced on a daily basis as opposed to an attempt to satisfy scientific curiosities. For example, ancient people were able to determine basic measurements such as length, distance, volume, weight, and time as a means to solve practical problems (Duncan, 1984). This assertion is backed by biblical references to the use of measurement and the writings of Aristotle mentioning civil officials checking weights and measures.

The first documented use of psychological testing dates back to 2,200 B.C. China where public officials were obligated to participate in civil service examinations every three years (DuBois, 1970). This competitive exam system assessed a variety of competencies such as archery, military affairs, agriculture, horsemanship, revenue, geography, music, writing, Confusion principles, knowledge of ceremonies, and civil laws. Examinees who scored well obtained appointments to governmental positions. Although these exams were rudimentary by modern standards, anecdotal evidence suggests a positive impact was produced, reducing the biases associated with nepotism and other political manipulations.

Darwin. Nevertheless, the roots of intelligence testing are embedded in the work of evolutionary theory and the use of systematic observation. According to Charles

Darwin's theory of evolution, the natural environment cannot support the reproductive capacity of organisms, leading to a struggle to survive (Darwin, 1859). Non-systematic mutations in the offspring of organisms result in variations or individual differences. These differences provide the offspring with adaptations which are more or less conducive to survival (Mader, 1996). As a result, when placed in a given environment, the characteristics that promote survival evolve through a natural process. Thus, over time, species undergo a slow transmutation whereby the characteristics that are associated with survival in a habitat occur with greater regularity.

Through his systematic observations of the variations across species, Darwin set in motion the development of scientific and statistical methods, producing a widespread impact on the field of modern psychology (Hergenhahn, 2009). For example, the roots of child and developmental psychology, comparative psychology, learning, abnormal psychology, testing and measurement, and, of course, evolutionary psychology can be traced directly to Darwin. In doing so, Darwin stimulated a curiosity in studying individual differences, raising questions regarding the link between human and animal intelligence.

Galton. The next major leap in the study of individual differences was advanced by Darwin's half-cousin, Francis Galton. Galton shared Darwin's infatuation with systematic observation (Clayes, 2001). In fact, Galton was so enamored with measurement that he attempted to measure a variety of phenomena such as the effectiveness of prayer (he did not find it effective), the degree of boredom at science lectures, and determine which country had the most beautiful women (Galton, 1883).

Galton is also credited with suggesting the use of fingerprints as personal identification, a practice later adopted by Scotland Yard (Forrest, 1974).

However, Galton's greatest advances to the field of measurement came when he opened the Anthropometric Laboratory in London's Health Exhibition. Visitors to the laboratory paid three or four pence each to have their sensory and motor abilities assessed, or for a smaller fee (two pence) an individual could be measured again at another time. In a little over a year, Galton collected measures on 9,337 subjects on variables such as height (standing), height (sitting), weight, arm span, lung capacity, pull strength, grip strength, keenness of sight, speed of blow (the time taken for someone to punch a pad), color discrimination, memory of form, hand steadiness, length of the middle finger, and auditory acuity (the ability to perceive or discriminate auditory tones) (Hergenhahn, 2009; Hothersall, 1995). Each individual received a copy of the results and Galton kept a copy on file (Irvine, 1986). Later, these data lead to the development of core statistical concepts such as correlation, regression to the mean, and the realization that as compared to mean (average) scores, median scores were less influenced by extreme scores (Bynum, 2002).

Galton was primarily interested in the inheritance of anatomical and cognitive abilities. According to Galton, intelligence was related to sensory acuity (Forrest, 1974). That is, the outside world is taken in through the senses. As such, individuals who possess keen senses were better able to acquire information. Since Galton believed that one's sensory acuity was directly related to intelligence, his laboratory is seen as the first effort to measure intelligence. However, Galton's contribution to the field of measurement is further realized through substantial methodological advances such as the development of the first assessment battery, a collection of sensory and motor measures. Likewise, many psychometric instruments in use today can be traced back to the work of Galton including rating scales, questionnaires, and self-report inventories (Hergenhahn, 2009).

Cattell. The popularization of psychological measurement in the United States is traced to the work of James McKeen Cattell. In addition to forming the first undergraduate psychology laboratory in the United States at the University of Pennsylvania, Cattell coined the term "mental test" (Cattell, 1890; Cattell, 1928; Hergenhahn, 2009). Moreover, Cattell is largely responsible for the encouragement of mental-testing research through his founding of several influential publications such as Psychology Review Science and American Men of Science. Likewise, Cattell founded the Psychological Corporation, which continues to be an industry leader in psychological testing and assessment.

From a methodological standpoint, Cattell introduced some critical assumptions about the validity of cognitive ability measures. For example, Cattell noted that if Galtonian measures were all measuring the same thing (i.e., intelligence), then they should all be highly correlated. Likewise, if a test is measuring intelligence, then it should demonstrate a substantial relationship with other indices of intelligence such as academic success. However, through his research, Cattell noted that Galtonian measures failed to demonstrate substantial relationships with one another or with other practical measures of intelligence such as college success (Guilford, 1967; Sternberg, 1990). As a result, sensory measures were deemed invalid indicators of intelligence and the interest in such measures faded.

Binet. Unlike Galton and Cattell who relied on sensory abilities as a proxy measure of intelligence, Alfred Binet proposed the study of mental abilities directly, by measuring higher mental processes through the use of variables such as memory, imagination, imagery, comprehension, attention, suggestibility, aesthetic judgment, force of will, moral judgment, and visual space judgment. Binet and Theodore Simon were commissioned by the French government to study children with mental retardation in French schools, culminating in the development of the Binet-Simon Scale of Intelligence, consisting of 30 tasks arranged in order of difficulty (Fancher, 1985). The measure was able to distinguish the performance of normal functioning and mentally delayed children, but later revisions also distinguished levels of intelligence in normal children, and provided normative information on adults (Siegler, 1992). Coupled with the addition of William Stern's coining of the term "mental age," a child's intelligence quotient (IQ) could be calculated as the child's mental age as derived from the Binet-Simon, divided by their chronological age (Fancher, 1985; Hergenhahn, 2009). The scale was again revised again by Lewis Terman, this time for American test takers, and validated against academic achievement ratings demonstrating the veracity of the test (Minton, 1988). The revised measure created by Terman is known as the Stanford-Binet Scale (Roid, 2005; White, 2000) and remains a measure of cognitive abilities.

Yerkes. The next major advance in cognitive assessment came as World War I dawned. The United States Army was faced with the problem of systematically evaluating and classifying the cognitive ability and emotional functioning of new soldiers (Boake, 2002). The influx of young men into the Army necessitated a method to quickly and efficiently assess and identify soldiers for selective training (e.g., officer training).

Robert Yerkes became familiar with intelligence testing while working at the Boston Psychopathic Hospital, suggesting a new scoring method to the Binet-Simon scale in which test takers are administered all items of the Binet-Simon scale, receiving credit (points) for the items passed (Hergenhahn, 2009). As such, intelligence could be measured by the items passed rather than by IQ, removing age as a factor and broadening the statistical analyses that could be performed leading to higher quality inference. However, the scoring and administration system devised by Yerkes had another benefit. Since the administration of the scale was not dependent on the age or ability level of the test taker, the items could be administered in a group setting.

When commissioned to develop an assessment device for soldiers, Yerkes maintained that such a test must measure innate intelligence and be easily administered and scored. The result was the Army Alpha, introduced in 1917, measuring verbal ability, knowledge or information, and ability to follow directions (Dahlstrom, 1985). A non-verbal equivalent version of the measure, the Army Beta, was introduced and administered to illiterate and non-English speaking soldiers. When testing was halted in 1919 following the end of the war, over 1.75 million people had been tested (Larson, 1994; McGuire, 1994). The success of the Army Alpha and Army Beta has led to the widespread use of group testing in schools and industry.

Cognitive Ability Testing in the Modern Era

Riding on the success of the Army Alpha and Army Beta, the use of paper-and-pencil cognitive ability measures has gained considerable popularity. The most commonly used measure of adult intelligence is the Wechsler Adult Intelligence Scale, currently in its fourth edition (WAIS-IV; Wechsler, Coalson, & Raiford, 2008).

Explicitly, Wechsler's tests are designed to measure "the global capacity of a person to act purposefully, to think rationally, and to deal effectively with his/her environment" (Wechsler, 1939, p. 229). As such, the WAIS-IV, which consists of 15 subtests, is primarily used in clinical settings due to the lengthy administration time, which is typically well over an hour. However, in occupational settings, where shorter measures are preferred, a comprehensive measure is neither required nor feasible (Chamorro-Premuzic, & Furnham, 2010).

Currently, the most widely used cognitive ability instrument in personnel assessment is the Wonderlic Personnel Test (WPT; Hunter, 1989; Wonderlic, 1992; Wonderlic, 2007). The WPT is a brief measure of cognitive ability that can be completed in approximately 15 minutes. Examinees are asked to answer as many of the 50 free-response verbal, quantitative, and spatial ability WPT items as possible within the allotted time limit (12 minutes). Despite its popularity, the relationship between WPT scores and intelligence is unclear. For instance, Bell, Matthews, Lassiter, and Leverett (2002) examined the relationship between WPT scores and the Kaufman Adult and Adolescent Intelligence Test (KAIT) finding the Wonderlic to be a robust predictor of both *Gf* and *Gc*. Conversely, Matthews and Lassiter (2007) conducted a similar study, examining the relationship between WPT scores and the Woodcock-Johnson-Revised (WJ-R), demonstrating that the Wonderlic is related to *Gc* but not *Gf*. As such, while WPT scores have been shown to reliably predict acquired knowledge, WPT have not been shown a reliable predictor of novel reasoning abilities. Likewise, the predictive power of WPT scores may be sample dependent whereas measures of *Gf* demonstrate robust relationships regardless of the administration samples (Hicks, Harrison, & Engle, 2015).

Another popular measure of cognitive ability, the Raven's Progressive Matrices Test (RPMT; Raven, Raven, & Court, 2003), is considered by some to be the best single measure of *Gf* and GMA available (Jensen, 1998; Nisbett et al., 2012). The RPMT is a "test of observation and clear thinking" (Raven, Court, & Raven, 1978, p. 3), requiring the examinee to inspect a matrix of geometric shapes linked by a common rule and extrapolate the next figure in the matrix that would satisfy the rule from several alternatives. Consisting of 60 items, the RPMT can be administered in 20 minutes and has been used extensively in the United States and the United Kingdom to make personnel selection decisions (Bertua et al., 2005; Jensen, 1998; Raven, Court, & Raven, 1998). Due to the non-verbal nature of the RPMT, it can be used across cultures without the need for item translations. As such, the terms *culture-free* (Cattell, 1940), *culture-fair* (Cattell & Cattell, 1963), and *culture-reduced* (Jensen, 1980) are all used to describe the Raven's and other similar non-verbal measures that require little cultural knowledge to answer test items. The advantage of culture-fair measures of cognitive ability is that they are thought to reduce the adverse impact seen in more culturally-loaded cognitive measures. Although culturally-fair tests have thus far not been shown to eliminate the adverse impact associated with measures of cognitive ability (Arvey & Faley, 1988), evidence suggests that reductions in adverse impact are obtainable using such measures over global intelligence measures (Hausdorf, LeBlanc, & Chawla, 2003).

Computer-Based Testing

Over the last 20 years, affordable, reliable, portable, and powerful computers have become a ubiquitous feature of our modern society, as seen in the omnipresence of desktop computers, laptops, tablets, and smartphones (Chernyshenko & Stark, 2015). Coupled with advances in online technology, computers offer an array of possibilities in the selection and presentation of assessment items, as well as where testing takes place (Sireci & Zenisky, 2006; Zenisky & Sireci, 2013). While early computerized assessments were little more than direct translations of paper-and-pencil measures (Barak & English, 2002), researchers are creating innovative computerized assessments that take advantage of the computing power afforded by such ubiquitous devices (e.g., Condon & Revelle, 2014). The technological revolution has led computerized assessment to rival the use of traditional pencil-and-paper methods as the dominate medium (Weiss, 2011).

No longer are assessments restricted by the limitations associated with traditional paper-and-pencil methods, such as static text statements and graphics. Rather, stimuli can be presented either audibly through computer speakers/headphones or graphically, moving through space on a computer monitor. The dynamic capabilities of computers allow for the creation and presentation of creative item formats previously unavailable to test developers. For example, three-dimensional computerized simulations and digital media are increasing the range of knowledge, skills, and other attributes that can be measured (Bartram, 2006; Jacobs & Chase, 1992; Zenisky & Sireci, 2002). Likewise, complex items that change over time can be created to improve the coverage of the constructs measured and their associated cognitive processes (Drasgow & Olson-Buchanan, 1999; Parshall & Harmes, 2009). For example, measures can be made of mouse or joystick movements or the time that elapses between item presentation and response, expanding the type and amount of information that can be obtained regarding test taker performance. As such, not only can the veracity of an examinee's response be called into question if only a few milliseconds elapsed between presentation of the item and the elicitation of a response (i.e., he or she did not read the question), but computerized testing allows for the complex scoring of the processes associated with producing a response (DiCerbo & Behrens, 2012).

Beyond the innovative item formats that are afforded by computerized testing, when coupled with the worldwide reach of the Internet, the benefits to organizations are staggering. For instance, organizations can reach a vast pool of potential applicants around the globe using a variety of measures without incurring the costs associated with printing and distributing measures via mail (Naglieri et al., 2004). As such, Internet-based test administration is more scalable and efficient than traditional pencil-and-paper measures all while presenting a consistent and positive image or culture to applicants of a company that uses advanced technology to staff employees (Tippins, 2009). Online measures also promote the standardization of measurement, uniformly presenting all test items in the exact same manner while improving the speed of processing applicants (Drasgow & Mattern, 2006; Drasgow & Olson-Buchanan, 1999; Tippins, 2009; Thurlow, Lazarus, Albus, & Hodgson, 2010; van der Linden & Glas, 2010). Due to these immense advantages, organizations see computer and Internet-based testing as an appealing alternative to traditional paper-and-pencil measures (Karim, Kaminsky, & Behrend, 2014).

Test Security Specific to Technology-Based Testing

Despite the advantages associated with computerized assessment, as with any technological advance, new and exploitable security threats arise. These security threats have caused some organizations to be hesitant to completely abandon the use of paper-and-pencil measures (Castella-Roca, Herrera-Joancomarti, & Dorca-Josa, 2006). Test security refers to a number of issues surrounding the test taker's ability to "cheat" or manipulate assessment scores through tactics such as possessing prior knowledge of the items, using outside sources, or using outsiders to answer test items (Karim et al., 2014). Online assessment is typically conducted in an unproctored testing environment, providing examinees a multitude of options to cheat, such as surfing the Internet or communicating with others to locate test answers (Al-Saleem & Ullah, 2014). Likewise, the proliferation of technological devices, such as smart phones, allow test takers to photograph, record, or otherwise document test content, and receive information virtually undetectably even under proctored conditions (Reynolds & Dickter, 2010). As such, although similar results are obtained from cognitive measures administered on computers and via traditional paper and pencil methods (Mead & Drasgow, 1993; Randall, Sireci, Li, & Kaira, 2012), practitioners and researchers have raised concerns that unproctored environments provide too great an opportunity to cheat. Consequently, test developers have warned against the administration of cognitive ability measures via online media (Naglieri et al., 2004).

Test security is a critical issue for test developers because it directly affects the validity of a measure (Foster, 2010). The use of impermissible sources or possessing prior knowledge of test items artificially inflates an examinee's score on the construct of interest. When compared to the scores of examinees that did not benefit from such a nefarious advantage, it erroneously appears that differing levels of the construct of interest are possessed. As such, any judgments or inferences based on compromised measures cannot be justified (Karim et al., 2014). Likewise, reductions in a test's validity directly affect its utility, which can have staggering financial implications for organizations (Schmidt & Hunter, 1998).

Security threats can be classified into six general categories of cheating (Foster, 2010). Threats include pre-exposing the test taker to test content, using a proxy to take the test, receiving help from someone at the exam center, using inappropriate aids during the assessment, hacking into the scoring database to raise or lower test scores, and copying the responses of another person during the exam. All of these methods represent an inappropriate or possibly illegal way in which test takers have attempted to inflate their assessment scores. However, none of these threats is limited to computer-based testing (Meyer & Zhu, 2013) and as such should not deter organizations from using such measures. Rather, these concerns shed light on designing and implementing improved methods to mitigate or eliminate such risks.

The most serious threat to exam security concerns test takers obtaining prior exposure to the test content (Foster, 2010). As compared to other threats to test security, prior knowledge of test content is often obtained inexpensively and with relative ease making it difficult to discriminate between honest and dishonest test takers. Moreover, the risk of being caught with prior knowledge of test content is extremely low. For example, an examinee may be provided information regarding the types of questions asked or specific item content and the associated correct answers prior to the administration. This form of cheating may be accomplished on computerized assessments by taking screenshots or otherwise documenting the items administered and then subsequently sharing the content with future test takers (Cook & Eignor, 1991). This problem is further compounded as testing windows become larger as is seen when organizations must continuously screen applicants (Croft, 2014). For instance, if thousands of examinees complete a measure comprised of the same items, the risk of sharing items increases greatly over time. This phenomenon was observed on a large scale by Asian students sitting for the Graduate Records Exam (GRE; Kyle, 2002). Examinees sitting for the exam at the beginning of the testing window copied exam content and shared it via online message boards. As a result, abnormally high scores were observed in the following months from countries such as China. Alarmed, GRE officials launched an investigation and uncovered websites containing exact test item content. As a result, the computerized version of the GRE was discontinued in the region, allowing only the paper-and-pencil version. Similarly, many organizations use only a single test form from which personnel decisions are based. As such, given a short measure, likeminded conspirators could memorize an entire scale in only a few administrations (Drasgow et al., 2009).

Combating Test-Security Issues

Traditionally, item sharing and other test security concerns have been combated by creating multiple forms of the same measure (Cook & Eignor, 1991). For example, 16 alternate forms of the WPT are available for use (Wonderlic, 1983). However, developing alternate forms is costly in both the time and financial resources required to generate them. As such, few alternate test forms are in use today (Freund & Holling, 2011). Moreover, while multiple forms may reduce an examinee's ability to memorize items from one administration to the next, developing parallel forms that are of similar content, difficulty, and reliability through traditional test development methods is virtually impossible, resulting in inequities across test forms (Cook & Eignor, 1991). Therefore, despite attempts to improve test security and fairness, test developers could inadvertently create a measure that is unfair in other respects.

Another method used to curb cheating is computer adaptive testing (CAT; Weiss, 2011). CAT creates a personalized test administration tailored to the examinee's ability level (Baylari & Montazer, 2009). Based on item response theory (IRT) methodology, CAT assumes that the test taker's ability level (i.e., amount of the latent trait) can be estimated by administering items of varying levels of difficulty. Examinees that possess greater ability levels of the latent trait are more likely to pass items of higher difficulty. Conversely, individuals with lower levels of the same trait may only pass items of lesser difficulty. Likewise, items that more finely discriminate a test taker's performance at a given ability level are said to provide more information at a given ability level since the ability level in question is tested more precisely. CAT takes advantage of IRT scaling by administering items of greater or lesser difficulty until the test taker's ability level can be estimated with an acceptable level of certainty (Babcock & Weiss, 2012; Weiss, 2011). Given that different items are administered to different examinees, some of the issues surrounding test security are addressed, but inequities in the items presented across test administrations may still exist. Likewise, advanced item exposure methods have sought to reduce test security concerns by controlling the frequency with which items are presented to specific geographic regions or time periods by adjusting a control parameter of an item exposure algorithm based on repetitive simulations (Chang & Ansley, 2003). However, CAT and item exposure algorithms do not prohibit test users from photographing or otherwise recording the items and distributing them to future test users. As such, given a large enough samples of test takers, test security in a CAT environment with exposure controls remains an issue necessitating the continual monitoring of item statistics to locate abnormal improvement in examinee performance (Drasgow & Mattern, 2006).

While the use of multiple test forms and CAT have been shown to reduce small-scale cheating, these methods require very large item pools (Drasgow et al., 2009). For instance, it is estimated that at least 2,000 items are needed to administer a 40-item CAT licensure exam twice a year (Breithaupt, Ariel, & Hare, 2009). As such, human test developers are strained to keep up with the demand for high quality items. Item generation is also a time consuming and costly process (Geerlings et al., 2011; Wainer, 2002). For instance, it is estimated that 10% of Educational Testing Service's (ETS) total testing costs are directly related to item writing (Wainer, 2002). Rudner (2009) suggested that development costs associated with the generation of a single item for a high-stakes licensure exam range from $1,500 to $2,000. As such, when Breithaupt et al.'s (2009) estimated number of items needed to create a 40-item item bank is combined with Rudner's (2009) cost-per-item estimate, the cost of the development of a high-stakes examination could reach $4,000,000.

However, not all items are created equal and the items generated by human test developers are often of questionable quality. The items created by human content specialists do not always conform to the construct of interest, nor can humans develop items that are of greater difficulty than they can conceived, placing a ceiling on the range of items that are possible (Hornke & Habon, 1986). The costs associated with test development are further increased as a substantial number of the items created by human developers must be eliminated from the item pool due to insufficient psychometric characteristics. For instance, Henryssen (1971) estimates that between 20 percent to 80 percent of the items generated by human test developers must be discarded during the test development process due to flaws. Thus, the use of automatic test development procedures has gained increased attention for the creation of cognitive ability measures, which are known to contribute to the prediction of occupational success (e.g., Schmidt & Hunter, 2004).

Automatic Item Generation

Given the need to quickly and efficiently generate large pools of items, automatic item generation (AIG) is a rapidly advancing field with roots in cognitive theory, computer technology, and psychometrics (Bejar et al., 2003). Also known as rule-based item construction, AIG is an alternative approach to traditional item development using computer technology to generate items based on item models and a set of rules (i.e., algorithm) that define item complexity (Gierl et al., 2015). The aim of AIG is to generate a large number of items that require little or no human review prior to administration (Doebler & Holling, 2015). Developing items in an AIG framework solves several of the practical issues associated with traditional test development. For instance, given an item model and a set of rules, AIG increases the flexibility of test administration through the generation of large pools of items of varying complexity with a negligible investment of time and money, reducing item exposure concerns (Geerlings et al., 2011). Additionally, since the items are generated through algorithms, precise information regarding how the items were constructed, their relation to the construct in question, and their psychometric properties is known (Geerlings et al., 2011). Moreover, the variety of AIG item types that can be created is ever expanding with research supporting their psychometric characteristics and test-retest applications (Arendasy & Sommer, 2013; Freund & Holling, 2011). As such, AIG is an attractive method for developing cognitive ability items (Freund, Hofer, & Holling, 2008; Poinstingl, 2009).

Under the AIG paradigm, item models (Bejar, 2002) serve as the basic structure upon which future items will be generated. Item models are either selected from exiting measures or uniquely created in such a way that the features of the model can be manipulated to create new items (Arendasy & Sommer, 2012; Gierl & Haladyna, 2012). That is, new items are generated from item models by specifying the construct-relevant features that can be varied, providing researchers a foundation for making inferences regarding test taker ability (Alves, Gierl, & Lai, 2010; Gierl et al., 2015).

Item model features known as "radicals" (Irvine, 2002), maximize the content-related variance in the items generated. That is, radical features define the processes or actions required to answer items. It is assumed that radicals systematically impact the psychometric characteristics (e.g., item difficulty) of items since they are selected based on the cognitive processes that test takers use to solve items. That is, radicals define the elements that are critical to solving an item and thus relate directly to item difficulty. Items that share radicals of the same complexity also share the same psychometric characteristics, such as measuring the same construct and item parameters (Doebler & Holling, 2015). Moreover, radicals can be varied independently of one another or used in tandem to generate an array of items that exhibit varying psychometric qualities (Arendasy & Sommer, 2012; Gierl et al., 2015). As such, researchers can create items of varying difficulty by manipulating of one or more radical elements. Thus, radicals improve the usefulness of the inferences that can be drawn from test taker performance since they allow for a widened range of the content domain to be tapped (Alves et al., 2010).

Unlike radicals, "incidentals" serve as the basis for generating variation in the surface features of items (Irvine, 2002) that do not directly relate to item difficulty. Incidentals do not exert an effect on the psychometric characteristics of an item, but rather change the "look" of items, creating variation within items of the same difficulty (Bejar et al., 2003). As such, the similarity among items with regard to psychometric characteristics is caused by radicals whereas item dissimilarity in terms of item appearance is caused by incidentals.

While it is a basic assumption of AIG that the effect of radicals affect test-taker performance in a similar way, this assumption may not hold in specific situations (Geerlings et al., 2011). That is, test takers may use different strategies to arrive at the same solution. In such circumstances, researches familiar with the cognitive processes used to answer items, as well as the radicals and incidents used to generate items, may not the potential for interference among the generative elements. As such, functional constraints (Arendasy et al., 2008) can be specified to omit certain combinations of radicals and/or incidentals that produce invalid test items, or items that interfere with the cognitive processes required to answer the question (Geerlings et al., 2012). For example, a researcher creating a mathematical ability measure may constrain the largest number that will be used in multiplication items to be less than 10 and omit all items that require the test taker to multiply by zero. These constrains not only serve as a quality control mechanism in AIG, but also avoid the generation of items that lead to solving items through the use of cognitive processes unrelated to the construct of interest and potential differential item functioning (Penfield & Camilli, 2007).

Moreover, since radicals exert a consistent effect on item difficulty, the effects of the radicals can be used to pre-calibrate items. As such, through the generation of items directly from previously calibrated item radicals and the random application of incidentals, items can be generated on the fly with predicable psychometric characteristics (Bejar et al., 2003). On-the-fly item generation is advantageous in that a large number of items are created in a fully automated fashion directly from calibrated radicals that define the item or item families (Geerlings et al., 2012). Moreover, test security concerns are lessened in that each test taker is provided a unique experience.

Item Model Development

As is the case in traditional item development, the expertise and creativity of content specialists is critical to designing meaningful AIG item models (Gierl, Lai, & Turner, 2012). Several published examples of the procedures and methods that researchers have used to generate item models exist (e.g., Arendasy & Sommer, 2012; Doebler & Holling, 2015; Freund et al., 2008; Geerlings et al., 2012; Gierl et al., 2015). However, despite the range of available tactics, AIG item modeling best practices is an under-researched area (Gierl & Lai, 2012).

As described by Arendasy and Sommer (2012), the number of useable items that are generated is related to the theoretical backing of the item model used in item construction. According to the authors, three AIG methods have been successfully used to generate cognitive ability items: (a) item modeling, (b) cognitive design system approach, and (c) automatic min-max approach.

Item modeling. Using the item-modeling approach, the researcher begins by selecting existing items from an operational measure. These parent items, also known as item models (Bejar, 2002), have radical features that can then be systematically varied to produce isomorphic iterations of the item. Due to their similarity to the parent item, the items created from this process are known as item clones (Glas & Van der Linden, 2003). Likewise, the item cloning process can be used to generate item sets or families of items that look different from one another, but are generated by the same combinations of radicals (Geerlings et al., 2011), resulting in item families that share similar psychometric characteristics. In theory, the newly created items would not need to be calibrated since their parameters can be drawn from known family distributions (Geerlings et al., 2011). Item modeling has been successfully used by ETS to supplement existing item pools. For example, Bejar (2002) developed a measure of quantitative ability through AIG item modeling methodology in which the researchers examined an existing pool of GRE quantitative items, choosing a subset of which to create item models.

The benefit of such an approach is that a test taker cannot simply memorize or solve the item by remembering an earlier solution (Gierl et al., 2015). For instance, a series of geometry items requiring the test taker to find the area of a rectangle could be created by simply changing the length of each side. Likewise, as indicated by Drasgow, Luecht, & Bennett (2006), item modeling, or the *weak theory* of item modeling, is well suited for a wide variety of content domains where few theoretical descriptions of the cognitive skills required in solving problems exist or unique item types are required. However, a drawback to this practice is that a large percentage of the items generated must be eliminated due to insufficient psychometric characteristics. Likewise, a relatively limited number of psychometrically distinct items can be created through the item modeling process. Since cloned items are vulnerable to the effects of test coaching (Morley, Bridgeman, & Lawless, 2004), and the ease with which examinees are able to recognize such items, the practice of item modeling is viewed negatively and regarded as overly simplistic (Gierl et al., 2015).

Cognitive design system approach. A more advanced approach to AIG relies on cognitive theory to guide item model construction. This *strong theory* of item model development (Irvine, 2002) begins with the examination and specification of the radicals that can be systematically varied on the basis of a cognitive model. As such, the level of difficulty resulting from the use of radicals can be predicted and subsequently tested to evaluate the contribution that the radical has to the prediction of item difficulties and to verify the use of the cognitive model. Subsequently, new item models are constructed to overcome the limitations exhibited by the current measure and the validity of the newly created item model is reexamined (Embretson, 2005). The use of cognitive theory and associated cognitive processes to guide decisions regarding which radicals will be manipulated as part of the item model is what differentiates this method from the *weak theory* item modeling approach.

The primary benefit of using a strong theoretical approach is the reduced need for extensive pilot testing since the factors that govern item difficulty can be specified, modeled, and controlled, allowing for the prediction of item difficulty (Gierl & Lai, 2012). Likewise, through the structured use of a cognitive model, item generation is enhanced through established empirical studies of cognitive functioning and individual-differences research. However, in practical applications, a considerable number of the items generated through the cognitive system design approach must be removed due to insufficient psychometric characteristics (Arendasy & Sommer, 2012). This issue is compounded due to the lack of available cognitive theories to guide item model development (Lai, Alves, & Gierl, 2009), limiting the use of the cognitive design system approach to narrow content domains such as mental rotation (Bejar, 1990) and abstract reasoning (Embretson, 2002). As such, similar to item modeling, researchers often resort to selecting items from existing measures to use as item models and constructing additional items that do not interfere with the other items, further restricting the number and quality of the items that can be generated (Arendasy & Sommer, 2012).

Automatic min-max approach. In order to overcome the limitations and loss of items resulting from insufficient psychometric characteristics associated with the item modeling and cognitive design system approaches, the automatic min-max approach was developed as a more sophisticated method of AIG which builds construct relatedness directly into the item construction process (Arendasy et al., 2008; Arendasy & Sommer, 2012). Compared to the cognitive design approach, the cognitive model specified in the automatic min-max approach initially covers a greater range of possible item formats, opening the possibly of a variety of innovative item types to tap the latent construct. As argued by Drasgow et al., (2006), AIG item modeling should be guided by the same design principles that are used in traditional test development (e.g., Downing & Haladyna, 2006). For example, the first step in traditional scale development is the clear statement of the latent construct that is intended to be measured from which future items can be written (Hinkin, 1998). Likewise, the first step in producing an item model through the top-down automatic min-max approach is a clear statement of the latent construct being measured along with the specification of the cognitive model that details the relevant knowledge, cognitive processes, and solution strategies that characterize the latent trait. Based on the cognitive model, the researcher then selects an item format to measure the latent trait. The cognitive model is then reduced to a more specific cognitive item model. This reduced model specifies the radicals that are thought to trigger the cognitive processes required to solve the item. Additionally, functional constraints are specified to omit specific item radicals and incidentals that may interfere with the cognitive processes of interest. As such, the automatic min-max approach is differentiated from the cognitive design approach through the use of a quality control mechanism and has been used to successfully generate algebra problems (Arendasy & Sommer, 2007), figural matrices (Arendasy & Sommer, 2005), mental rotation (Arendasy & Sommer, 2010), number series (Arendasy & Sommer, 2012), and English and German word-fluency (Arendasy, Sommer, & Mayr, 2012) items with little to no loss in items due to insufficient psychometric characteristics.

Procedural Framework of the Automatic Generation of Analogical Reasoning Problems

As detailed by Arendasy and Sommer (2012), the automatic min-max approach to AIG-model development includes the specification of the latent trait under consideration; choice of item format; specification of the cognitive model; and specification of the radicals, functional constraints, and incidentals. Likewise, items of the type discussed in the following section require the generation of alternative answer choices. This section describes the procedural framework of item model construction for the experimental AIG measure used in this study.

Definition of the Latent Trait

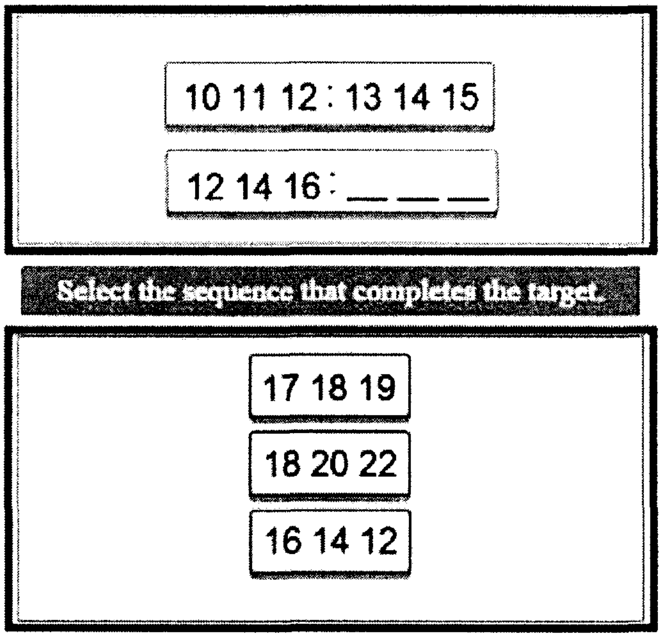
Previous research has indicated that *Gf* is closely related to *g* and is characterized by the ability to solve novel problems and adapt to new situations (Cattell, 1957, 1971; Gustafsson, 1984, 1989, 2001; Schneider & McGrew, 2012). Measures that best capture *Gf* are relatively culture-free, non-verbal, spatial measures of inductive reasoning (Carroll, 2003; Sattler, 2001). To clarify, inductive reasoning is the ability to identify trends or patterns and extrapolate this information to reach a logical conclusion (Raven, 1938). In contrast, deductive reasoning is the ability to apply one or more given rules to obtain a solution (Shye, 1988). Therefore, inductive reasoning entails the discovery of relationships while deductive reasoning does not. As modeled by Spearman (1938), inductive-reasoning items are solved by examining the elements of a problem, determining the logical relationships between them, and extrapolating these relationships to other elements. As such, the abilities associated with inductive reasoning are typically measured by tests consisting of analogies, classifications, matrices, and series (Goldman & Pellegrino, 1984; Sternberg & Gardner, 1983; van de Vijver, 1991).

Choice of the Item Format

In order to capture the latent trait and capitalize on AIG methodology, the researcher chose a unique analogical item type. Analogical reasoning is the ability to draw relationships between objects in one context and use this information to explain the same relationship in another context (French, 2002; Holyoak, 2005). As such, the substantial cognitive component of tasks such as these is the integration of multiple complex relationships (Robin & Holyoak, 1995). Closely related to Raven's-type tasks (Snow, Kyllonen, & Marshalek, 1984), analogical reasoning items require the examinee to describe, generalize, or explain new phenomena based on familiar concepts, and serves as a basis for dealing with novelty. Thus, the ability to reason through analogy is critical for everyday situations and is closely linked to *Gf* (Duncan et al., 2000; Holyoak & Morrison, 2005; Prabhakaran, Smith, Desmond, Glover, & Gabrieli, 1997).

In the current study, the experimental AIG measure of *Gf* was assessed through the use of analogical reasoning number sets (See Figure 1). The choice of the item type was not taken lightly. In order to create a measure as devoid of cultural influences as possible, numbers were chosen as a medium due to their near universal use (Porter, 1995). While numbers are used to represent values or quantities, it can be argued that other symbols (e.g., letters, arrows, shapes) may impart unintended representations depending on the cultural lens from which they are viewed (Bradley, 2010). As compared to other symbols, numbers provide a means to assess examinee *Gf* abilities through symbols that are familiar to most cultures.

In the experimental AIG measure, examinees were presented with a series of automatically-generated number sets consisting of three numbers in an A:B::C:D (A is to B as C is to D) sequence. Specifically, a randomly generated number set is presented in Term A. In Term B, the number set is transformed according to one or more "rules." The examinee's task is to identify the rule(s) that govern the number set transformation from Term A to Term B. The examinee is then asked to apply the previously identified rule(s) to another randomly generated number set in Term C to obtain the number set that would occupy Term D from three multiple-choice alternatives.



10 11 12 : 13 14 15

12 14 16 : __ __ __

Select the sequence that completes the target.

17 18 19

18 20 22

16 14 12

Figure 1. Sample Experimental AIG Measure Item

Specification of the Cognitive Item Model

The cognitive processes that are involved in solving analogical reasoning items can be arranged into a series of stages (Evans, 1968; Mulholland, Pellegrino, & Glaser, 1980; Sternberg, 1977). Although various models and terms have been used to describe the cognitive processes associated with answering analogical reasoning items (Mulholland et al., 1980), Sternberg's (1977), cognitive process and naming conventions are used below. In the first stage, *Encoding*, a mental representation of the individual terms of the analogy are created, allowing further mental operations to be performed. In the second stage, *Inferring*, the relationship between the corresponding attributes of first two terms (A – B) is inferred and stored in working memory. In the third stage, *Mapping*, the relationship between the first and third terms (A and C) is discovered and, likewise, is stored in working memory. In the fourth stage, *Application*, the relationships discovered in the Inferring and Mapping stages are used to identify the correct answer for the fourth term (D). In an optional stage, *Justification*, particularly used in answering True-False analogical reasoning items (e.g., Mulholland et al., 1980), the previous stages are checked to determine if an error is made or to determine if additional information is required to answer the question. In the final stage, *Response*, an answer is physically selected or marked on an answer sheet from response alternatives.

Specification of Radicals, Functional Constraints, and Incidentals

The automatic min-max approach to AIG item model development requires the formal specification of the radicals, functional constraints, and incidentals that promote content representation within the items generated (Arendasy & Sommer, 2012). Each of these elements is described in this section.

Radicals. As previously described, radicals define the processes or actions required to answer items. As such, radicals relate to the difficulty of the items generated. Primi (2001) describes the complexity factors that influence the difficulty of *Gf* items. These complexity factors are analogous to item radicals in AIG methodology. Primi details four complexity factors: the number of elements, the number of rules, the types of rules, and the perceptual organization of items. The "number of elements" refers to the number of attributes contained in an item, while the "number of rules" refers to the number of radical elements that are invoked by a given item. Both of these factors are associated with the cognitive load that is placed on the operational capacity of working memory (Mulholland et al., 1980; Salthouse, 1994). As noted by Carpenter, Just, and Shell (1990), participants completing matrices items decompose the items into smaller sub-goals, requiring participants to track an increased number of elements in order to satisfy higher goals. Thus, as additional attributes and rules are applied to items, strain is placed on the limited capacity of working memory. The "types of rules" refers to the complexity of the content that is applied to the item attributes. For example, Jacobs and Vandeventer (1972) created a taxonomy of the transformations that can be used to manipulate figural matrices items. These transformations range from simple rules (e.g., changes in object size) to complex transformations (e.g., adding matrices attribute) that influence item difficulty. However, as noted by the authors, in practice, matrices items do not cover the content domain well in that certain transformations tend to be overrepresented or oversampled. Finally, "perceptual organization" refers to the visual complexity or esthetics of the items. As described by Primi (2001), "visually harmonious items display perceptual and conceptual combinations that represent congruent relationships between elements, whereas nonharmonic items tend to portray competitive or conflicting combinations between visual and conceptual aspects that must be dealt with in reaching a solution" (p. 51). For example, Carpenter et al. (1990) noted that misleading cues such as superposed elements in matrices items increase item complexity. Likewise, Primi (2001) demonstrated that over 50% of the variance in item complexity is accounted for by perceptual organization.

Number of elements. The experimental AIG measure was designed to allow for the lengths of the number sets used to be variable. However, for practical purposes, the number sets in the current study were restricted to three numbers. Since the number of elements included in each item is consistent, the length of the number sets is not expected to exert a cognitive load on examinees.

Number of rules. The number sets used in the experimental AIG measure were generated according to a set of rules described in the next section. Since it is possible to generate items that result from the application of one or more rules, additional cognitive load is expected to be exerted as additional rules are applied to the analogical numbers sets.

Type of rules. The type of rule applied to the number sets should also influence item difficulty. In order to link and manipulate the terms of the experimental AIG measure, mathematical operations were applied to the number sets. Namely, consistent mathematical operations and mixed mathematical operations were used as radicals. Consistent mathematical operations included problems in which the examinee was presented with a randomly generated number and then addition was applied to obtain the second number in the sequence; the third number was then obtained by again applying addition to the second number (e.g., 15 – 16 – 17 : 22 – 23 – 24). The same consistent mathematical operation applied if subtraction was used to obtain the second number from the first, and the third from the second. Conversely, mixed mathematical operations consisted of items in which addition (or subtraction) was applied to the first number to obtain the second, and then the opposite mathematical operation subtraction (or addition) was applied to the second number to obtain the third (e.g., 14 – 16 – 12 : 20 – 22 – 18). Likewise, numbers within the number series could duplicate. As such, the duplication of numbers will also be used as a radical.

Perceptual organization. As noted by Primi (2001), the perceptual organization of item stimuli substantially impacts item difficulty. As such, the visual complexity of the automatically-generated number sets is expected to influence the overall difficulty of the items. For example, number sets that maintain the same perceptual organization across terms (e.g., 7 – 8 – 9 : 12 – 13 – 14) are expected to be less difficult than items in which the perceptual organization of the items is flipped between terms (e.g., 5 – 6 – 7 : 14 – 13 – 12). As such, the visual dissimilarly of the numbers within a number sets should influence item difficulty.

Functional constraints. Functional constraints are specified to minimize the influence of unintended cognitive processes in solving AIG items. The goal of functional constrains is to enhance the construct relatedness of the AIG items created such that the abilities other than that of the construct of interest are removed from the item model. As such, the AIG items created conform more closely to the intended construct.

Based on the item type and cognitive model, the constraints placed on the AIG item model can take many forms. For instance, the random numbers contained in the number sets will be constrained to two digits (10-30) to control the cognitive complexity of the items generated (Horn & Noll, 1997). Likewise, ambiguous items that permit more than one solution should be prohibited (Scharroo & Leeuwenberg, 2000). That is, it is conceivable that AIG items could be generated in which a correct solution and a distractor number series are identical. In this case, a comparison can be made between answer choice alternatives. If two answers are identical, another item can be generated. Similarly, studies of analogical reasoning tasks demonstrate the effect of stimulus priming on task performance (Blanchette & Dunbar, 2002; Spellman, Holyoak, &

Morrison, 2001; Wharton, Holyoak, & Lange, 1996). Therefore, the radicals presented to the test takers should be randomized to mitigate the effects of pre-exposure of identical radicals.

Incidentals. Incidentals are designed to create variation in item appearances, but have no effect on item difficulty. In the current study, item variation is achieved by randomly generating numbers to populate the number sets.

Distractor Generation

As in traditional test development, AIG item difficulty is dependent on producing distractors that are plausible enough to be endorsed (Doebler, 2015; Downing & Yudkowsky, 2009). For example, in developing a static multiple-choice measure, incorrect options that are endorsed by at least 5% percent of test takers are considered "functional distractors" while response options that are endorsed to a lesser degree add little value to a measure (Downing & Yudkowsky, 2009). In AIG, algorithms are used to create distractors (Gierl et al., 2012). However, simplistic strategies such as randomly selecting items from the universe of options will result in items that are too easy since the correct option is easily identified (Doebler, 2015). Rather, the psychometric soundness of AIG measures can be improved by systematically switching or removing the radicals or combinations of radicals that were used to generate the item stem (Arendasy & Sommer, 2005; Doebler, 2015). For example, if the numbers in Term C are linked by adding *2*, a distractor item may fail to add *2* or add a number other than *2* to generate an incorrect option for Term D. As such, controlled variation is produced in the response options, masking the correct answer and improving the measurement of the construct of interest. While the procedures just mentioned represent the current best practices in AIG, the effects that distractors have on the psychometric properties of items is difficult to ascertain and is an under represented area of researcher (Gierl et al., 2012).

Formulation of the Problem

The purpose of the current research is to build on the existing AIG methodological framework through the construction and validation of an on-the-fly measure of cognitive ability that is generated at the time of item presentation. In order to fulfill this purpose, the proposed measure will be developed using the automatic min-max approach (Arendasy & Sommer, 2012). Next, the psychometric characteristics and the nomological network of the experimental AIG measure will be examined. The general expectations are that the proposed measure will demonstrate unidimensionality and construct relatedness and will correlate with other measures of *Gf.*

A fundamental concern in the development of a psychological instrument is the establishment of the unidimensionality of the measure. Dimensionality refers to the number of latent traits that contribute to responding to the items of an instrument (DeVellis, 2012). Commonly, the dimensionality of psychometric instruments is evaluated through the use of exploratory and confirmatory factor analysis. However, as recommended by Arendasy and Sommer (2012), the unidimensionality of AIG measures can be assessed through the use of the Rasch model as a prerequisite for testing the constructed relatedness of AIG items. The fit of the data to the Rasch is examined through the use of likelihood ratio tests (e.g., Andersen, 1973; Martin-Löf, 1973), which relate the likelihood of the item parameter data to a null model. If the tests fail to reach significance, then the hypothesis that the experimental AIG measure demonstrates Rasch model fit can be retained. As such, Hypothesis 1 concerns the dimensionality of the experimental AIG measure. It is expected that the experimental AIG measure will display unidimensionality.

Hypothesis 1: The experimental AIG measure will display unidimensionality.

Construct representation (Embretson, 1983) concerns the identification of the theoretical operations that contribute to performance on a measure. For AIG measures, the construct representativeness of a measure is determined by examining the effects that the specified item radicals contribute to item difficulty (Embretson & Daniel, 2008; Freund et al., 2008; Gierl & Haladyna, 2012; Poinstingl, 2009). As such, construct representation provides evidence supporting the inclusion of the item radicals in the item model since these elements are hypothesized to affect item difficulty. Thus, initial evidence for the construct representation of the experimental AIG measure is established through the examination of these features (Arendasy & Sommer, 2012). The ultimate goal is to produce a model that accounts for as much item difficulty as possible, based on the features of the item model (Gierl & Haladyna, 2012). As such, Hypothesis 2 concerns the content representation of the experimental AIG measure. It is expected that the item radicals specified will predict item difficulty.

Hypothesis 2: The experimental AIG measure will display satisfactory construct representation.

Hypothesis 2a: Consistent Mathematical Operations will significantly predict item difficulty.

Hypothesis 2b: Mixed Mathematical Operations will significantly predict item difficulty.

Hypothesis 2c: Duplicate Numbers will significantly predict item difficulty.

Hypothesis 2d: Flipped Relationships will significantly predict item difficulty.

In traditional test-development applications, test-retest reliability is commonly used to demonstrate the stability of test scores across administrations (Anastasi & Urbina, 1997). As noted by Shuttleworth (2009), measures of cognitive ability are good candidates for such analyses because it is unlikely that participant ability level will suddenly change. Thus, it is expected that participants will obtain similar scores across test administrations. Therefore, Hypothesis 3 concerns the temporal stability of the experimental AGI measure.

Hypothesis 3: The experimental AIG measure will show adequate test-retest reliability.

Another method used to demonstrate the validity of a measure is to examine its nomological network (Cronbach & Meehl, 1955). As described by Campbell and Fiske (1959), convergent validity provides an indication that a measure shares a substantial relationship to other measures to which it should be theoretically related. As noted previously, non-verbal and culture-free measures of inductive reasoning best capture *Gf* (Carroll, 2003; Sattler, 2001). As conceived by Thurstone, tasks such as Letter Sets and Number Series tap inductive reasoning abilities well (Freedheim & Weiner, 2003). When subjected to confirmatory factor analysis, along with matrices measures, a substantial *Gf* factor is formed by Letter Sets and Number Series tasks (Hicks et al., 2015). The purpose of this series of analysis is to examine the criterion relatedness of the experimental AIG measure. As such, Hypotheses 4a and 4b concern the predictive relationship shared between the experimental AIG measure and established measures of cognitive ability. It is expected that the experimental AIG measure of *Gf* will correlate with other measures of *Gf*.

However, *Gf* is also known to share a relationship with demographic variables. For instance, previous research indicates that *Gf* tends to decrease with age (Cattell, 1943). Therefore, Hypothesis 4c concerns the predictive relationship shared between the experimental AIG measure and examinee age. It is expected that the experimental AIG measure of *Gf* will correlate negatively with examinee age.

Hypothesis 4: The experimental AIG measure will demonstrate satisfactory criterion validity.

Hypothesis 4a: The experimental AIG measure will significantly predict scores on the Letter Sets task.

Hypothesis 4b: The experimental AIG measure will significantly predict scores on the Number Series task.

Hypothesis 4c: The experimental AIG measure will demonstrate a significant negative relationship with examinee age.

CHAPTER 2

DEVELOPMENT OF THE EXPERIMENTAL AIG MEASURE

The development of the experimental AIG measure began with a content analysis. The purpose of the content analysis was to identify the item radicals, incidentals, and functional constraints that could be manipulated and controlled. Four content specialists who hold advanced degrees in psychological sciences served as subject matter experts (SMEs) in this analysis. SMEs were provided with a definition of the latent trait, the cognitive model, and a prototypical item model. SMEs were then asked to examine the item model and verbally describe the process an examinee would take to solve a given item. Likewise, SMEs were asked to indicate the various elements of the item model that could be varied in order to trigger the appropriate solution strategy. Using the information provided by the SMEs, radicals, incidentals, and functional constraints were specified.

Generative Matrix

Based on the information obtained from the content analysis, the experimental AIG measure was created using the PHP programing language, a popular open source server-side scripting language (PHP.net, 2016). In order to generate the analogical reasoning items, first, a randomly-generated base number was produced for each of the four analogical reasoning terms (A through D) and multiple-choice alternatives. Base number values were constrained to numbers between 10 and 30 in order to reduce the cognitive load associated with interpreting number values greater than two digits in length and to add perceptual uniformity to the look of AIG items. Likewise, this constraint served to eliminate the possibility of negative values. These randomly-generated numbers were intended to serve as incidental elements, creating variation in how the items look without affecting difficulty. Next, term manipulation numbers were randomly generated for use in subsequent mathematical manipulations. These term manipulation numbers were used to create patterns in the analogical reasoning terms. Term manipulation numbers were constrained to values between 1 and 4 in order to limit the cognitive load associated with adding and subtracting numbers of lower or higher values. The term manipulation values of Terms A and B were identical as were the values for Terms C and D. For instance, if the number *4* was generated to manipulate Term A, *4* was also used to manipulate Term B. Likewise, if the number *3* was generated to manipulate Term C, *3* was also used to manipulate Term D.

Next, mathematical and logical manipulations were applied to the base numbers using the term manipulation numbers as controlled by an item generation matrix. The item generation matrix consisted of 14 variables (See Table 1) dictating item and distractor construction. The leftmost column represents the item being generated. The 14 columns to the right (labeled 1 through 14) represent the variables manipulated to generate the items. In the table, each variable is listed below the aspect of the item that is controlled. Further clarification on how the items are generated is presented in the following paragraphs.

The mathematical manipulation between the first and second value of each term was controlled by Variable 1. Variable 1 could take one of three values (1 = subtraction; 2 = addition; 3 = duplicate). For example, suppose the base number for Term A was *10* with a term manipulation number of *3*. The value of Variable 1 dictates if 3 is added or subtracted from 10. If Variable 1 had a value of *2*, 3 was added to 10 to generate the second value (13) in Term A. Likewise, if the value of Variable 1 was *1*, then 3 would be subtracted from 10 to generate the second value (7) in Term A. However, if instead the value of Variable 1 was *3*, then the term manipulation number (3) would be ignored and 10 would be a duplicated value (10). Variable 2 acted in the same manner, controlling the relationship between the second and third number in the term.

Table 1. *Item Generation Matrix*

	Item Gen		FL	FR	Dis1		Dis2		Dis3		FD1	FD2	FD3	#Ds
Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1	1	1	1	1	1	2	2	1	2	1	1	2	2
2	1	2	1	1	1	2	2	1	1	3	1	2	2	2
3	1	3	1	1	1	3	3	1	2	1	1	1	2	2
4	2	1	1	1	2	1	1	2	2	2	1	2	2	2
5	2	2	1	1	2	2	1	1	2	3	1	1	2	2
6	2	3	1	1	2	3	3	2	3	1	1	2	2	2
7	3	1	1	1	3	1	1	3	3	2	1	1	2	2
8	3	2	1	1	3	2	2	3	1	1	1	2	2	2
9	1	1	2	1	1	1	2	2	1	2	1	1	2	2
10	1	2	2	1	1	2	2	1	1	3	1	2	2	2
11	2	1	2	1	2	1	1	2	2	2	1	2	2	2
12	2	2	2	1	2	2	1	1	2	3	1	1	2	2
13	3	1	2	1	3	1	1	3	3	2	1	1	2	2
14	3	2	2	1	3	2	2	3	1	1	1	2	2	2
15	1	2	1	2	1	2	2	1	1	3	2	2	1	2
16	2	1	1	2	2	1	1	2	2	2	2	2	1	2
17	3	1	1	2	3	1	1	3	3	2	2	1	1	2
18	3	2	1	2	3	2	2	3	1	1	2	2	1	2
19	1	2	2	2	1	2	2	1	1	3	2	2	1	2
20	2	1	2	2	2	1	1	2	2	2	2	2	1	2
21	3	1	2	2	3	1	1	3	3	2	2	1	1	2
22	3	2	2	2	3	2	2	3	1	1	2	2	1	2

Note. Item Gen = item generation; FL = flip left; FR = flip right; Dis1 = distractor 1; Dis 2 = distractor 2; Dis3 = distractor 3; FD1 = flip distractor 1; FD2 = flip distractor 2; FD3 = flip distractor 3; #Ds = number of distractors.

In addition to Variables 1 and 2 that generate differentiation in the pattern of numbers in each term, Variables 3 and 4 were used to influence the perceptual complexity of analogical reasoning items. These variables allowed for the pattern created by Variables 1 and 2 to be "flipped," expanding the construct space, requiring the test taker to examine and draw relationships across item terms. For example, a term consisting of the numbers 15 – 16 – 17 could be inverted to 17 – 16 – 15 if indicated by Variables 3 or 4. Variables 3 and 4 could take one of two values (1 = no flip; 2 = flip) with Variable 3 controlling Terms A and C and Variable 4 controlling Terms B and D.

The remaining variables in the item matrix were used to generate item distractors. Variables 5 through 10 are analogous to Variables 1 and 2, controlling the generation of the pattern of numbers that comprise the three distractor terms. Variables 5 and 6 controlled Distractor 1; Variables 7 and 8 controlled Distractor 2; and Variables 9 and 10 controlled Distractor 3. The patterns of variable values were systematically manipulated to create plausible distractor choices. For example, if Variables 1 and 2 contained values of *1* and *2* respectively, Variables 5 and 6 may consist of values *1* and *2*, *2* and *1*, *1* and *1*, or *2* and *2*. Additional variation in distractor items was produced by systematically manipulating the term manipulation number. For example, if the term manipulation number was *3*, distractor items may contain values surrounding this value (e.g., 1, 2, or 4). Variables 11, 12, and 13 are analogous to Variables 3 and 4 controlling the "flip" of the terms in Distractors 1, 2, and 3 respectively. Thus, the "flip" variables applied to the distractors allow for additional variation and further mask the identification of the correct answer. Additionally, constraints were placed on distractor terms eliminating the possibility that a distractor matched the correct answer. Finally, Variable 14 was used to indicate the number of distractors to generate. Variable 14 could take on values of 1, 2, or 3 indicating how many distractors to generate. In the current study, this variable was held constant at 2, allowing for the presentation of only two distractors and a correct answer. However, future research may examine the effects of greater or fewer distractors.

The values of the 14 variables in the item generation matrix were counterbalanced to create uniform variation and broad construct coverage in the 22 AIG analogical reasoning item families and the multiple-choice distractors. Moreover, since item generation was controlled using an item matrix, precise information about how the items were constructed allowed for the precise testing of the radicals and incidental involved.

A 22-item measure comprised of the items generated from the item generation matrix was administered online along with a demographic form which asked basic information including age, gender, ethnicity, and educational attainment. Consistent with scale development best practices (DeVellis, 2012; Freund & Holling, 2011), a demonstration of the rules (i.e., radicals) that were used to manipulate analogical reasoning terms was presented via an instructional video. Additionally, written instructions were made available to participants. As noted by Freund et al. (2008), tests of inductive reasoning frequently suffer from a lack of clarity regarding the types of tasks that are involved in solving items. For instance, the rules that govern the relationship between corresponding analogical reasoning terms must be discovered before the relationship discovered can be extrapolated (Sternberg, 1977). However, with no additional information, it is plausible that test takers may disagree on the rules that govern the relationship between terms. Thus, unintended rules may be applied to items that allow a test taker to reach a solution that is quite different from the "correct" solution. In order to avoid this issue, test takers can be presented information regarding the various rules that govern the relationship between analogy terms prior to test administration.

In addition to clarifying the task, test fairness and accuracy of the measure is increased since no participant is unfairly disadvantaged by misunderstanding the patterns imbedded in the items (Freund et al., 2008). As such, before the measure is administered, practice items were made available to participants, allowing them to become familiar with the task and item format that was used. Participants were allowed to complete as many practice items as they wished without time constraints. During the practice session, participants were provided feedback regarding the correctness of each response.

After completing the practice session, the presentation of the item radicals used to generate the 22-item experimental measure was randomized to control for order effects. Once participants selected a response, they were not able to return to the previous item.

In order to reduce examinee fatigue and to limit the amount of time taken to complete the experimental AIG measure, a pilot study was conducted to determine the amount of time provided to answer each item. Participants ($N = 4$) were asked to answer the items of the experimental AIG measure as quickly as possible. The mean response time was 16.51 seconds ($SD = 8.45$). As a result of the pilot study, a 30-second time limit was established for examinees to answer each item.

Scoring

Raw scores for the experimental AIG measure were calculated using the following scheme. First, the average response time for items that were answered correctly was calculated from the total sample ($M = 15.06$ seconds). Due to its approximation of the midpoint of the time allowed to answer the items, this figure was rounded down to 15.00 seconds, and this served as a benchmark value. Next, a score of *1* was awarded to participants who answered the item correctly and submitted their response prior to the benchmark value. A *0* was awarded for items that were either answered incorrectly and/or elicited a response after the benchmark value. Since the experimental AIG measure contained 22 items, scores could range from 0 to 22.

CHAPTER 3: STUDY 1

CONTENT VALIDATION RESULTS AND DISCUSSION

The purpose of the current research is to build on the existing AIG methodological framework through the construction and validation of an on-the-fly measure of cognitive ability that is generated at the time of item presentation. In order to accomplish the aims of the research, three studies were conducted examining the construct representation, temporal stability, and criterion relatedness of the scores produced by the experimental AIG measure. The aim of Study 1 is to assess Hypotheses 1 and 2 relating to the unidimensionality and construct representation of the experimental AIG measure.

Participants

The sample consisted of 333 respondents (193 male and 140 female) from the United States between the ages of 18 and 81 (*M* = 35.3; *SD* = 13.8). Guidelines for traditional scale development suggest that a sample of approximately 300 participants is required to ensure the stability of the findings (Nunnally, 1978). Likewise, Downing (2003) indicates that a sample of at least 200 participants is required to assess Rasch model fit. As such, the size of the sample in the current study seems adequate.

The majority of the participants (80.8%) were recruited through Amazon's Mechanical Turk (mTurk). mTurk has become a popular crowdsourcing platform from which behavioral science researchers may solicit research participants (Chandler, Mueller, & Paolacci, 2014; Krupnikov & Levine, 2014). Previous research has indicated that the results obtained from mTurk workers are comparable to conventional sources of data collection such as convenience and snowball sampling (Buhrmester, Kwang, & Gosling, 2011; Casler, Bickel, & Hackett, 2013). Likewise, previous scale-development initiatives have sourced mTurk workers as participants, producing scales with acceptable psychometric characteristic (Buhrmester, et al., 2011). The remaining participants (19.2%) were recruited via snowball sampling through social media. In exchange for their participation, participants sourced from mTurk were provided monetary compensation. Prior to data collection, a pilot study was conducted to estimate the average amount of time required to complete the experimental AIG measure. mTurk workers were compensated according to this time estimate and to the median minimum wage for the United States to ensure fair wage compensation. All participants were provided feedback regarding their performance on the AIG measure (number of items answered correctly).

Of the sample, 76.6 % identified as White/Caucasian, 9.3% as African-American, 6.9% as Hispanic-American, 2.1% as Asian-American, 0.6% as American-Indian, and 8.4% as other. The reported educational attainment levels were as follows: 0.3% some school, no high school diploma; 13.2% high school diploma or equivalent; 18.9% some college credit; no degree; 3.9% trade/technical/vocational training; 10.8% Associate degree; 36.9% Bachelor's degree; 11.1% Master's degree; 1.5% Professional degree;

2.7% Doctorate degree. The average time spent working on the test was 8:31 minutes (*SD* = 3:18 minutes) ranging between 2:21 and 25:27 minutes.

Procedure

Study data were collected via an online measure hosted by the researcher. In order to access the scale, participants were provided a link to the experimental AIG measure. Before beginning the measure, participants were presented with an informed-consent form stating the purpose of the project, instruments involved, risks and alternative treatments, compensation (if any), and the contact information of the researcher. The letter of approval from the Louisiana Tech University institutional review board (IRB) is presented in the appendix. Participants were then asked to provide basic demographic information (age, gender, race, and educational attainment). Participants were instructed to answer items as quickly as possible and were given the opportunity to complete as many practice items as they wished. Practice items were administered without time limitations, and feedback regarding the correctness of responses was provided after an answer was selected. Once comfortable with the task, participants could advance to the actual experimental AIG measure. Participants were provided 30 seconds in which to respond to each item. If an answer was not submitted in less than 30 seconds, participants were automatically advanced to the next question. No feedback was given regarding the correctness of items in the non-practice portion of the measure.

Results

Analyses were conducted using SAS 9.1, SPSS 17.0 (descriptive and correlational values), and RStudio (LLTM analysis). Prior to performing the analyses, item responses that were submitted in less than one second were recoded as missing data. These suspect responses were likely the result of participants inadvertently double clicking the response button to the previous question. Since the presentation of the items was randomized, these suspect items can be classified as missing completely at random (Little & Rubin, 2002). Maximum likelihood estimation was used to impute the missing data points. Previous research has indicated that maximum likelihood estimation is advantageous to other methods of handling missing data including listwise and pairwise deletion, as well as mean imputation techniques (Baraldi & Enders, 2010; Newman, 2003). Of the total sample, 70 of the cases required the imputation of one or more items. Subsequent analysis of imputed and non-imputed cases revealed that the scores of these measures correlated highly ($r = .99$). On average, participants answered 17.15 ($SD = 4.30$) of the 22 items correctly within the 30 seconds provided for each item. However, once the item scoring algorithm was applied, the mean score obtained on the measure was 9.20 ($SD = 5.12$). A one-sample Kolmogorov-Smirnov ($p = .001$) and Shapiro-Wilk test ($p < .001$) indicate that participant total score data were not normally distributed. However, after a visual inspection of a histogram plot (Figure 2), it was determined that the dataset exhibited sufficient normality (Howell, 2013).

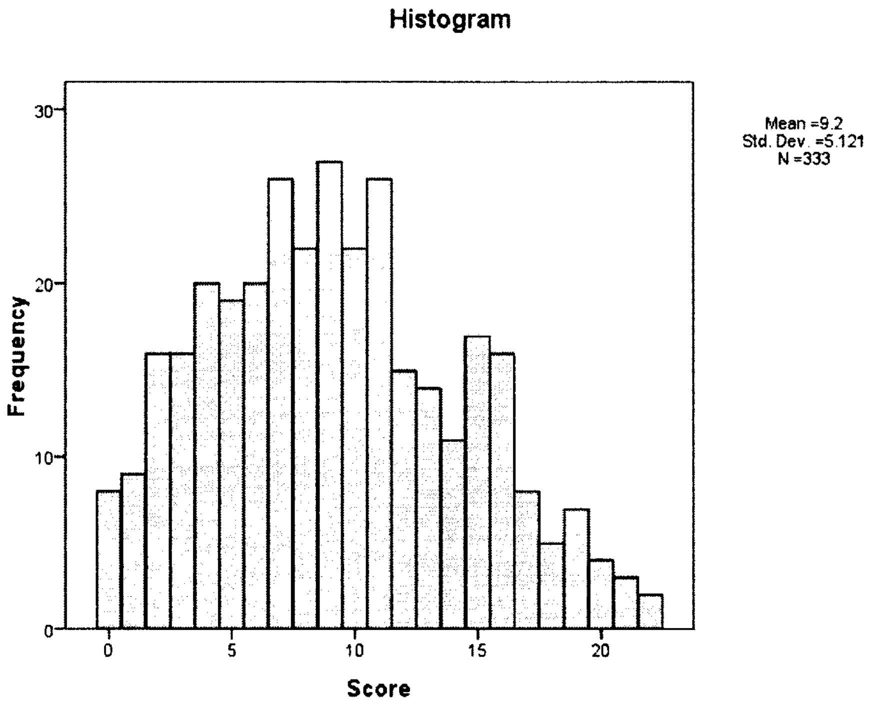


Figure 2. Histogram of Experimental AIG Scores

Previous research examining gender differences in *Gf* reveal no systematic differences (Colom & García-López, 2002). This is not to say that males and females perform equally well on all *Gf* tasks. For example, meta-analytic evidence indicates that in adult samples, males tend to outperform females (*d* = .33) on tasks such as Raven's Advanced Progressive Matrices (Lynn & Irwing, 2004). In general, females tend to outperform males on verbal tasks while males outperform females on spatial-ability measures (Halpern, 1997; Neisser et al., 1996). Thus, when measures of verbal, reasoning, and spatial ability are combined to obtain broad ability estimates, gender gaps are largely eliminated. However, as reported by Casey, Nuttall, Pezaris, and Benbow (1995), males are particularly advantaged in mathematical-ability tasks. This difference is principally seen at the upper end of the ability continuum. In contrast, no gender differences in mathematical ability are seen in low-ability and average-ability samples. As noted by Brody (1992), the difference in mathematical ability may be due to highly developed visual-spatial skills in such high-ability males.

The experimental AIG measure tasks the examinee with quickly identifying mathematical manipulations and drawing relationships across a visual-spatial field. Therefore, one may expect to see differential scoring on such a measure. An independent-samples *t*-test was conducted to compare gender differences in scoring on the experimental AIG measure. Results of the analysis indicate a significant effect for gender, $t(326.11) = 3.69, p < .001$, with men receiving higher scores than women. Likewise, an independent-samples *t*-test was conducted to compare score differences between the mTurk and snowball samples. Results of the analysis indicate a significant sample effect, $t(331) = -2.74, p = .007$, with the mTurk sample receiving higher scores than the snowball sample. For examinees who provided demographics, the scores produced by male and females at six age intervals for the total, mTurk, and snowball samples are presented in Tables 2, 3, and 4 respectively. Likewise, the scores produced by males and females by educational attainment are presented in Table 5.

Table 2. *Experimental AIG Measure Score Means and Standard Deviations by Gender for Six Age Intervals.*

	Male			Female			Total Sample		
Age intervals	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
18-29	63	11.0	5.6	52	8.1	3.9	115	9.7	5.1
30-39	75	10.3	5.2	37	8.0	5.1	112	9.6	5.2
40-49	24	9.0	4.5	22	7.3	3.7	46	8.2	4.2
50-59	15	8.9	5.3	15	9.5	4.4	30	9.2	4.8
60-69	7	5.9	6.4	4	10.3	7.9	11	7.5	6.9
70+	6	7.7	8.0	4	3.5	1.3	10	6.0	6.4
Total	190	10.0	5.5	134	8.0	4.4	324	9.2	5.2

Table 3. *Experimental AIG Measure Score Means and Standard Deviations by Gender for Six Age Intervals for the mTurk sample.*

	Male			Female			Total Sample		
Age intervals	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
18-29	59	11.0	5.8	40	8.3	4.0	99	9.9	5.3
30-39	65	10.3	5.2	28	8.3	5.2	93	9.7	5.2
40-49	22	9.1	4.7	21	7.2	3.8	43	8.2	4.3
50-59	15	8.9	5.3	11	9.9	4.8	26	9.3	5.0
60-69	2	14.0	5.7	2	13.0	9.9	4	13.5	6.6
70+	2	16.5	.7	0	n/a	n/a	2	16.5	.7
Total	165	10.4	5.3	102	8.3	4.5	267	9.6	5.1

Table 4. *Experimental AIG Measure Score Means and Standard Deviations by Gender for Six Age Intervals for the snowball sample.*

	Male			Female			Total Sample		
Age intervals	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
18-29	4	11.3	4.0	12	7.7	3.7	16	8.6	3.9
30-39	10	10.3	5.6	9	7.0	5.0	19	8.7	5.4
40-49	2	7.5	.7	1	8.0	n/a	3	7.7	.6
50-59	0	n/a	n/a	4	8.5	3.8	4	8.5	3.8
60-69	5	2.6	2.6	2	7.5	7.8	7	4.0	4.5
70+	4	3.3	5.3	4	3.5	1.3	8	3.4	3.5
Total	25	7.6	5.7	32	7.1	4.1	57	7.3	4.8

Table 5. *Experimental AIG Measure Score Means and Standard Deviations by Gender for Eight Educational Attainments.*

	Male			Female			Total Sample		
Educational attainment	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
High school grad	27	9.5	5.0	17	7.1	5.9	44	8.6	5.4
Some college	33	11.2	5.8	30	7.4	3.8	63	9.4	5.2
Trade/technical	5	8.4	4.7	8	5.6	3.0	13	6.7	3.8
Associate's degree	24	9.5	5.5	12	8.7	4.8	36	9.3	5.3
Bachelor's degree	74	9.5	5.6	49	8.4	3.9	123	9.1	5.0
Master's degree	18	10.4	4.2	19	9.9	4.6	37	10.2	4.3
Professional degree	4	12.8	5.1	1	4.0	n/a	5	11.0	5.9
Doctorate degree	6	11.3	6.8	3	4.0	4.4	9	8.9	6.9
Total	191	10.0	5.4	139	8.0	4.4	330	9.1	5.1

Some classical test theory results. Using the precedent set by Doebler and Holling (2015), classical test theory analyses commonly reported in scale development research are presented here to aid in the interpretation of the psychometric characteristics of the experimental AIG measure. These statistics are meant to provide the reader with a more complete understanding of how the measure is performing. In general, Cronbach alpha values of .70 or greater indicate acceptable internal consistency (Kline, 1999). However, Kline also notes that cognitive ability measures should strive for alphas of .80 or greater. The experimental AIG measure demonstrated adequate internal consistency (α = .86; *SEM* = 1.91). Likewise, Nunnally and Bernstein (1994) suggest that item discrimination values of greater than .20 are sufficient while Anastasi and Urbina (1997) propose that item difficulty values between .15 and .85 are acceptable. Item discrimination values ranged from .34 to .52 (median = .43) while difficulty values ranged from .14 to .63 (median = .46) indicating that the item statistics largely conform to recommended tolerances.

Linear Logistic Test Model (LLTM). The evaluation of item radicals and incidentals can be accomplished either through the LLTM (De Boeck & Wilson, 2004; Fischer, 1973, 1995) or multiple regression analyses (Gorin & Embretson, 2006). In AIG studies, the LLTM is more commonly employed since it provides a means to evaluate cognitive models (Arendasy et al., 2008; Arendasy & Sommer, 2010, 2012; Arendasy et al., 2012; Freund et al., 2008). That is, LLTM allows for the empirical testing of the cognitive processes that contribute to item difficulty, thus demonstrating construct validity of the items generated from the item model (Fischer, 1973). Under the LLTM, the difficulty parameter of the Rasch model is reduced into a linear combination of radical effects, allowing for their contribution to the prediction of item difficulty to be assessed (Freund et al., 2008; Holling, Bertling, & Zeuch, 2009). That is, the LLTM assumes that the difficulty parameter of the Rasch model is comprised of several cognitive operations that sum to the overall difficulty parameter estimate (Baghaei & Kubinger, 2015). As such, there is no point in decomposing the difficulty parameter of a Rasch model that lacks fit, as the data produced would lack meaning (Fischer, 2005). Therefore, assessing the fit of the Rasch model is prerequisite for applying the LLTM (Fischer, 1973; Poinstingl, 2009).

The consistency of the Rasch model can be assessed through likelihood ratio tests determining the fit of the data to the model. As noted by Rost (1982), the Rasch model assumes both item and person homogeneity. As such, both forms of homogeneity must be tested. Tests of item homogeneity determine if more than one person parameter is needed. Conversely, checking for person homogeneity entails determining if more than one item parameter is needed for each item to describe the data.

The Martin-Löf (1973) test for unidimensionality is a likelihood ratio test used to examine the fit of the Rasch model by separating the items of a measure into two groups of items. The item parameters of these groups of items are subsequently examined for homogeneity (Mair, Hatzinger, & Maier, 2013). If the maximum likelihood values of both sets of items are approximately equal to the maximum likelihood calculated for both sets of items together, then the Rasch model holds, and it is assumed that both sets of items tap the same dimensions (Verguts & De Boeck, 2000). Thus, non-significant values indicate that Rasch model holds. The Martin-Löf results failed to reveal a significant difference (median raw score: χ^2 [120] = 82.36, $p > .05$).

The Andersen (1973) likelihood ratio test was also used to determine the fit of the data to the Rasch model. This test compares the item parameters of two predefined subgroups in the total sample to determine if differential item functioning is present as a result of the splitting criterion (Futschek, 2014). In AIG studies, median raw scores are commonly used as the partitioning criterion (Freund et al., 2008; Arendasy & Sommer, 2012). If the likelihood ratio test fails to reach significance, then the fit of the data to Rasch model is retained and the LLTM can be estimated (Baghaei & Kubinger, 2015). The results of the Andersen test indicate that the data fit the Rasch model (median raw score: χ^2 [21] = 24.79, $p > .05$). As such, Hypothesis 1 concerning the unidimensionality of the AIG measure is supported.

Under the LLTM, item difficulty is calculated based on the weighted contribution of the item radicals through a design matrix, indicating the degree to which these elements are related to the cognitive complexity of AIG items (Embretson & Daniel, 2008). As a result, the combined effects of radicals can be used to account for the difficulty parameter in the Rasch model, supporting the construct representation of the item model (Arendasy & Sommer, 2012). The hypothesized cognitive components that are required to solve assessment items are entered as a Q-matrix. The Q-matrix used in the current analysis is presented in Table 6. The columns of the Q-matrix represent the cognitive operations measured by the experimental AIG measure, and the column values indicate the weights applied to each of the cognitive process for each item. For instance, the number series pattern in Item 1 (e.g., 10 – 8 – 6 : 17 – 15 – 13) consisted of subtraction between the first and second number, and subtraction between the second and third number (Consistent Mathematical Operation). However, Item 2 (e.g., 10 – 8 – 12 : 17 – 15 – 19) consisted of subtraction between the first and second number and addition between the second and third numbers (Mixed Mathematical Operations). As such, the Q-matrix details the hypothetical cognitive components (i.e., radicals) that are thought to influence item difficulty.

Table 6. *Q-Matrix for the Experimental AIG Measure*

Item	Consistent Mathematical Operations	Mixed Mathematical Operations	Duplicate Numbers	Flipped Relationship
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	1	0	0
5	1	0	0	0
6	0	0	1	0
7	0	0	1	0
8	0	0	1	0
9	1	0	0	1
10	0	1	0	1
11	0	1	0	1
12	1	0	0	1
13	0	0	1	1
14	0	0	1	1
15	0	1	0	1
16	0	1	0	1
17	0	0	1	1
18	0	0	1	1
19	0	1	0	0
20	0	1	0	0
21	0	0	1	0
22	0	0	1	0

Radical difficulties are assessed through an easiness parameter (i.e., eta). The easiness parameters, standard errors, and 95% confidence intervals for each of the four hypothesized radicals in the LLTM analysis are presented in Table 7. Negative easiness parameter values indicate cognitive operations that increase the difficulty of items while positive values indicate radicals that can reduce the difficulty of items (Baghaei & Kubinger, 2015). As such, the item radicals of "Mixed Mathematical Operations" and "Flipped Relationships" increase the difficulty of items, while "Consistent Mathematical

Operations" and "Duplicate Numbers" decrease the difficulty of items. As suggested by Baghaei and Kubinger (2015), if the confidence intervals that surround the easiness parameters fail to include zero, the radical specified significantly contributes to item difficulty. Radicals that fail to support the predicted relationship with item difficulty can be excluded from the item generation process and the cognitive model can be redefined. All radicals of the experimental AIG measure significantly predict item difficulty. As such, Hypotheses 2a through 2d regarding the ability of the radicals to significantly predict item difficulty is supported.

Table 7. *Parameter Estimates, Standard Error, and Confidence Intervals for the Item Radicals of the Experimental AIG Measure*

Parameter	Estimate	*SE*	Lower CI	Upper CI
Consistent Mathematical Operations	0.500	0.171	0.166	0.834
Mixed Mathematical Operations	-1.050	0.133	-1.310	-0.790
Duplicate Numbers	0.550	0.177	0.204	0.896
Flipped Relationships	-0.703	0.058	-0.815	-0.590

Further demonstration of the construct representation of the LLTM analysis is indicated by the correlation of the empirically generated Rasch easiness parameters to those predicted by the LLTM analysis. As indicated by Arendasy and Sommer (2013), R^2 values of .70 and greater are desirable. The empirically derived and predicted item difficulty parameters of the experimental AIG measure were highly correlated (r = .97). Thus, the R^2 value in this analysis was .93, indicating that 93% of the variance in the Rasch difficulty parameter could be accounted for by the cognitive model. The plot of the empirically derived and predicted item difficulty parameters is presented in Figure 3. As such, Hypothesis 2 regarding the construct representation of the radicals of the experimental AIG measure is supported.

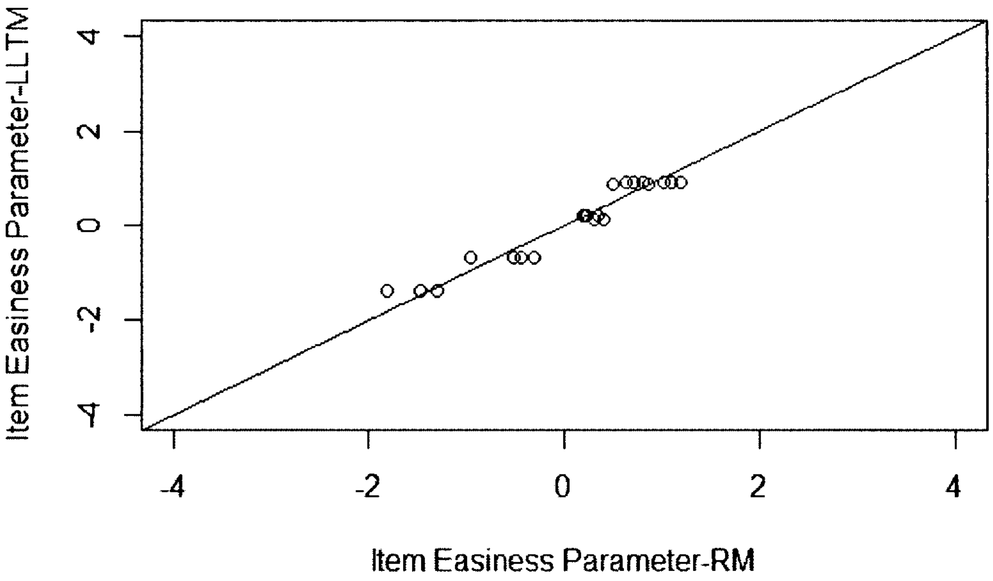


Figure 3. Plot of Empirically and LLTM Generated Item Difficulty Parameter Estimates

Discussion

The purpose of the current study was to understand the construct representation of an experimental AIG measure by assessing the influence of the hypothesized cognitive components on item difficulty. Conventional item analysis suggests that although the items of the AIG measure were generated from a random base number and the presentation of radicals was randomized, the measure demonstrates adequate internal consistency (.86) for a measure of mental abilities. Likewise, median item discrimination (.43) and difficulty (.46) values produced are within established guidelines. From a classical test construction perspective, the items of the experimental AIG measure are well correlated, able to distinguish test taker performance, and of appropriate difficulty.

The results of Study 1 also show that it is possible to model the experimental AIG measure data through Rasch and LLTM, allowing for an estimation of the contribution of the influence that radicals impart on item difficulty. According to the Martin-Löf and Andersen tests, item and person homogeneity are present, supporting unidimensionality of the experimental AIG Measure and fit to the Rasch model.

Using LLTM, Mixed Mathematical Operations had the greatest impact on increasing the difficulty of items followed by Flipped Relationships. Presumably, each of these radicals placed a cognitive load on working memory reducing the likelihood of obtaining a correct answer within the time allotted. In contrast, Consistent Mathematical Operations and the inclusion of a duplicate number within a term had the opposite effect, lowering the difficulty of the items generated. As such, this result provides evidence supporting the inclusion of the hypothesized item radicals in the item model since these elements affect item difficulty. Likewise, the hypothesized cognitive model accounted for a large portion (93%) of the variance in the Rasch difficulty parameter, producing results that are similar to other LLTM investigations (e.g., Arendasy, 2000, 2005; Arendasy & Sommer, 2005, 2007; Arendasy et al., 2007; Gittler, 1990; Gittler & Arendasy, 2003). As such, the cognitive model specified demonstrates substantial coverage of the processes test takers use to obtain a correct response to the items of the measure. Thus, the analyses detailed in Study 1 support the assertion that experimental AIG measure demonstrates adequate unidimensionality and construct representation.

CHAPTER 4: STUDY 2

TEMPORAL STABILITY RESULTS AND DISCUSSION

Study 2 was designed to assess the temporal stability of the experimental AIG measure across test administrations. Although this type of analysis is not commonly performed on on-the-fly AIG measures, test-retest correlations are commonly used in classical test design to describe scale functioning. As such, this study is designed to provide insights into the stability of experimental AIG measure scores over time.

Participants

A subset of Study 1 examinees elected to participate in Study 2. The sample consisted of 36 respondents (22 male and 14 female) from the United States between the ages of 21 and 71 ($M = 37.69$; $SD = 14.59$). According to Field (2009), samples of this size ($N \geq 30$) are generally sufficient for research purposes. The majority of the participants were recruited through mTurk (77.8%). The remaining 22.2% of participants were recruited via snowball sampling through social media. Of the sample, 66.7% identified as White/Caucasian, 11.1 % as African-American, 8.3% as Hispanic-American, 5.6% as Asian-American, 2.8% as American-Indian, and 11.1% as other. Likewise, participant educational attainment levels were as follows: 16.7% high school diploma or equivalent (e.g., GED); 8.3% some college credit, no degree; 2.8%

trade/technical/vocational training; 8.3% Associate degree; 50.0% Bachelor's degree; 11.1% Master's degree; 2.8% Doctorate degree.

Procedure

Experimental AIG measure data were collected on two occasions. Participant data for the first administration was collected as part of Study 1. A subset of the participants who completed Study 1 was invited to complete the measure for a second time. Approximately one week following the first administration, participants were provided with the link to the experimental AIG measure for a second time and asked to complete the scale. For each administration, total scores were calculated using the same scoring scheme described in Chapter 2. A total of 47 participants completed the experimental AIG measure twice. Due to suspected changes in the manner in which examinees approached the second administration, examinees who obtained score differences greater than 3 SEMs across administrations were removed from the test-retest sample. As a result, 11 people were removed from the sample to arrive at the total sample ($N = 36$). The mean number of days between administrations was 8.78 ($SD = 2.38$).

Results

The current analysis tests the temporal stability of the experiential AIG measure by assessing the reliability of the measure over two testing sessions. As noted by McCrae, Kurtz, Yamagata, and Terracciano (2011), test-retest reliability is conceptually independent of internal consistency, reflecting the consistency of scores obtained on separate occasions. Anastasi and Urbina (1997) state that test-retest reliability "shows the extent to which scores on a test can be generalized over different occasions; the higher the reliability, the less susceptible the scores are to random daily changes in the conditions of the examinee or the testing environment" (p. 92). In general, test-retest values of .70 or greater are considered acceptable (Andrews, Peter, & Teesson, 1994; Burlingame, Lambert, Reisinger, Neff, & Mosier, 1995). The administration means, standard deviations, and test-retest correlation are presented in Table 8. As shown, the correlation between two experimental AIG measure administrations is acceptable. Therefore, the results of this study support Hypothesis 3 and the temporal stability of the experimental AIG measure.

Table 8. *Test-Retest Reliability for the Experimental AIG Measure*

	First Testing		Second Testing		
Experimental AIG Measure	M	*SD*	M	*SD*	*r*
Total Score	9.50	4.14	9.89	4.96	.80*

Note. *N* = 36; **p* < .001.

Discussion

Study 2 was designed to assess the relationship that the experimental AIG measure shares with itself across test administrations. The results of this analysis indicate that the experimental AIG measure correlates well with itself (.80). Previous test-retest research using AIG measures has obtained similar results (Freund & Holling, 2011). However, to the researcher's knowledge, this is the first test-retest study of an on-the-fly AIG measure. Rather, previous research used static items created using AIG methodology, mimicking traditional test-retest methods. As such, these measures are susceptible to the same practice effects seen in paper-and-pencil measures of cognitive ability. Given that each participant was administered assessments consisting of different items at an average interval of slightly over one week, the results obtained from the current analysis are promising. Although a higher test-retest value is desirable, the items of the experimental AIG measure may contain item features (e.g., radicals and incidentals) that we have yet to identify or control. Likewise, modified scoring schemes allowing for partial credit may improve the temporal consistency of the scores obtained from the measure. As such, supplemental research examining the manipulation of the basic item model and score calculations may produce more robust test-retest figures.

CHAPTER 5: STUDY 3

SCALE VALIDATION RESULTS AND DISCUSSION

Study 3 was designed to assess the relationship between the experimental AIG measure and other measures of *Gf*. To this end, the experimental AIG measure was correlated with two established measures of *Gf*. Likewise, the relationship between the AIG measure and age, which is known to be related to *Gf*, was examined. Thus, the aim of Study 3 is to assess Hypotheses 3 concerning the criterion validity of the experimental AIG measure.

Participants

A subset of Study 1 examinees elected to participate in Study 3. The sample consisted of 31 respondents (12 male and 19 female) from the United States between the ages of 19 and 81 (*M* = 43.76; *SD* = 17.58). According to the central limit theorem, samples of greater than 30 participants will approximate a normal distribution (Field, 2009). As such, the size of the sample in the current study is adequate. All participants were recruited via snowball sampling through social media. Of the sample, 74.1% identified as White/Caucasian, 6.5% as Hispanic-American, and 19.4 % as other. Likewise, participant educational attainment levels were as follows: 9.7% some college credit, no degree; 3.2% trade/technical/vocational training; 38.7% Bachelor's degree; 32.2% Master's degree; 16.1% Doctorate degree.

Procedure

Study data were collected using three measures (described below). Participant data on the experimental AIG measure were collected as part of Study 1. A subset of the participants who completed Study 1 was invited to complete the criterion validation measures. Following the completion of the experimental AIG measure, participants were provided with a unique identifying code and redirected to a survey containing the criterion validation measures hosted on Qualtrics.com. Before beginning the criterion validation measures, participants were instructed to enter a unique identifying code allowing the scores obtained from the experimental AIG measure and validation measures to be linked.

Measures

AIG Measure. The independent measure of *Gf* was assessed using the same experimental AIG measure used in Study 1. The researcher invited a subset of the Study 1 participants to participant in the current analysis after completing the 22 item experimental AIG measure. Total scores were based on the scoring procedure described in Chapter 2.

Letter Sets. Letter Sets (Set 1) (Ekstrom, French, Harman, & Dermen, 1976) measures an examinee's ability to identify patterns in groups of letters and was used as a measure *Gf*. Each item consists of five four-letter strings (e.g., NLIK, PLIK, QLIK, THIK, VLIK). The examinee's task was to identify the rule shared by four of the five strings and eliminate the string that does not conform to the rule. Seven minutes were provided to complete the 15-item measure. Scores range from 0-15 with higher scores indicating better performance. Previous research indicates that Letter Sets are relatively culture-free measures, independent of quantitative or verbal abilities, provide an efficient measure of *Gf* and require only a minimal investment of time (Duran, Powers, & Swinton, 1987). Redick, Unsworth, Kelly, & Engle (2012) estimate the internal consistency of Letter Sets to be .78. Likewise, when subjected to confirmatory factor analysis, the Letter Sets task loads substantially (.81) on the *Gf* factor, indicating appreciable fit to the construct (Hicks et al., 2015).

Number Series. Number Series (Thurstone & Thurstone, 1962) measures mathematical-inductive reasoning, and is thought to be primarily influenced by *Gf* (Kvist & Gustafsson, 2008). Each item of the measure consists of a series of numbers (e.g., 10, 11, 12, 13, 14). The examinee's task is to identify the underlying mathematical relationship shared between terms and extrapolate the next number in the sequence. Examinees have 4.5 minutes to complete the 15-item measure. In a longitudinal study, Schaie (2005) reports that the Number Series task displays a test-retest reliability of r = .77 and a seven-year test retest reliability r = .74. Likewise, Kvist and Gustafsson (2008) report that the Number Series task loads substantially (.81) on the *Gf* factor when subjected to confirmatory factor analysis, suggesting a strong fit to the construct.

Results

In this analysis, the correlational relationships between the experimental AIG measure and established measures of *Gf* are presented. Since this analysis assesses the theoretical relationship between the experimental measure and criterion measures, it is necessary to correct for a lack of reliability in the criterion (Letter Sets and Number

Series), but not the independent measure (experimental AIG measure) (Ghiselli, Campbell, & Zedeck, 1981; Guilford, 1954; Guion & Highouse, 2006; Schmitt & Klimoski, 1991). Failure to correct for unreliability artificially weakens coefficient values and masks the true relationship (Salgado, Moscoso, & Anderson, 2016). The means, standard deviations, and corrected and uncorrected correlations for the experimental AIG and criterion measures are presented in Table 9. As suggested by Hopkins (2002), the following guidelines can be used to interpret the correlations: coefficients between .00 and .09 are very small or trivial; coefficients between .10 and .29 are small; coefficients between .30 and .49 are moderate; coefficients between .50 and .69 are large; coefficients between .70 and .89 are very large; and coefficients between .90 and 1.00 are nearly perfect. Using Hopkins's conventions, the correlations between the experimental AIG measure and the criterion measures in Table 9 are classified as "large." As such, the results of this analysis support Hypotheses 4a and 4b as indicated by a significant relationship between the experimental AIG measure and criterion measures of Letters Sets and Number Series, respectively.

Table 9. *Means, Standard Deviations, and Correlations Between the Experimental AIG Measure and Criterion Measures*

Measure	*M*	*SD*	1	2	3
1. Experimental AIG Measure	7.58	4.48	--		
2. Letter Sets	9.61	3.35	.50 (.44*)	--	
3. Number Series	7.71	2.87	.61 (.54**)	.90 (.70***)	--

Note. $n = 31$; Corrected correlation coefficients are outside of parentheses; Uncorrected correlation coefficients inside of parentheses; *$p < .05$; ** $p < .01$; ***$p < .001$

Previous research has also noted that *Gf* is related to demographic variables. For instance, Cattell (1943) suggests that *Gf* tends to decrease with age. Therefore, the scores of the experimental AIG measure should decrease as a function of examinee age. Using the participant data described in Study 1, the means, standard deviations, and corrected and uncorrected correlations for the experimental AIG measure and age are presented in Table 10. The result of this analysis supports Hypothesis 4c as indicated by a significant negative relationship between the experimental AIG measure and examinee age.

Table 10. *Means, Standard Deviations, and Correlations Between the Experimental AIG Measure and Age*

Measure	*M*	*SD*	1	2
1. Experimental AIG Measure	9.20	5.12	--	
2. Age	36.30	12.67	-0.16**	--

Note. *N* = 333. **p* < .05; ***p* < 01; ****p* < .001

Discussion

Study 3 was designed to assess the relationship that the experimental AIG measure shares with established measures of *Gf*. After correcting for unreliability in the criterion measures, the scores obtained from the experimental AIG measure and Letter Sets (Ekstrom et al., 1976) produced a correlation of .50. Likewise, using the same correction, the scores of the experimental AIG measure and Number Series (Thurstone & Thurstone, 1962) produced a correlation of .61. Using Hopkins's (2002) evaluative guidelines, these correlations are described as "large." As such, the results of this study indicate that the experimental AIG measure is tapping the construct of *Gf* as measured by other established measures.

The negative relationship between age and *Gf* has been noted for some time. As described by Cattell (1943), the nature of *Gf* is such that ability gains are seen through adolescence and then diminishes with age. Past research examining the longitudinal relationship between *Gf* and perceptual speed support the generalized slowing of processing abilities as one ages (Schaie, 1989). Likewise, Bors and Forrin (1995) found that the relationship between age and *Gf* was reduced to a nonsignificant value after controlling for mental speed, indicating that the decrement of *Gf* with age is substantially related to processing speed. These findings are buttressed by findings linking *Gf* and shorter reaction times (Grabner, Fink, Stipacek, Neuper, & Neubauer, 2004). The experimental AIG measure was designed as a brief measure of *Gf*, forcing examinees to respond quickly to items. As such, it is not surprising that in the current study, a significant negative relationship was found between the experimental AIG measure and age. This result provides limited support of the assertion that the experimental measure is tapping aspects of *Gf*.

While the results obtained from the current study are promising, it should be noted that the measures used in this study consisted of relatively brief criterion measures of *Gf*. McGrew (2009) notes that *Gf* is associated with myriad of inductive and deductive tasks. As such, future studies should be conducted on a diverse array of *Gf* instruments to better understand the relationship that the experimental AIG measure has with *Gf*. However, taken together, the results of this study largely support the assertion that the experimental AIG measure conforms to the *Gf* construct, particularly as measured by established criterion measures.

CHAPTER 6

DISCUSSION

In the psychological sciences, perhaps no construct has received as much attention as cognitive ability. Although competing perspectives and theoretical orientations have emerged regarding the nature of intellectual functioning, there is overwhelming evidence that generalized intelligence plays a key role. Across situations *g* demonstrates a predictable influence on academic success (Ones et al., 2006), workplace performance (Schmidt & Hunter, 1998), and problem solving in everyday situations (Gottfredson, 2002). Due to the substantial relationship shared with *g*, *Gf* is regarded as the backbone of intellect (Arendasy et al., 2008; Gustafsson, 1984, 1989, 2001). Consisting of the ability to adapt to new situations and solve novel problems (Cattell, 1957, 1971; Gustafsson, 1984, 1989, 2001; Schneider & McGrew, 2012), *Gf* is best measured by non-verbal and culture-free tasks such as number series and analogical reasoning problems (Sattler, 2001).

Despite a long psychometric tradition associated with the measurement of cognitive abilities, researchers have embraced technological advancements as a means to uncover what it means to be smart. For instance, computer technology has provided test developers with a dynamic platform to present an immense array of unique test items types. Computers can display graphical figures and images of greater complexity than could be conveyed through paper-and-pencil administration. Coupled with internet technology, assessments can be delivered to all corners of the globe in a cost effective and convenient manner. However, such unproctored administrations raise the issue of test security, such as cheating and item sharing (Cook & Eignor, 1991), limiting the acceptance of the results produced (Naglieri et al., 2004).

Historically, test developers have used multiple test forms or CAT administrations to combat test security issues. While these methods are able to curtail some of the threats to test security as compared to fixed measures (Guo, Tay, & Drasgow, 2009), these methods also require large pools of continuously updated, psychometrically sound items. However, it has become clear that costly and inefficient methods of traditional item construction by human item writers cannot keep pace with the growing demand. Likewise, the items created by such means often lack the psychometric rigor needed to seed item pools. As such, researchers have begun to explore advanced methods to generate high-quality test items.

Rooted in cognitive and computer sciences, AIG methodology allows researchers to specify the structural elements that define item difficulty to produce large pools of items with known psychometric characteristics (Geerlings et al., 2011). The rapidly advancing field of AIG methodology has gained a considerable amount of attention from the psychometrics community for its ability to quickly, efficiently, and cost effectively produce vast pools of items based solely on an item model and a computer algorithm (Gierl et al., 2015). In doing so, the AIG framework solves many of the practical issues and threats to test security that have hindered test construction and administration.

In the present study, a unique item type designed to measure *Gf* was developed using AIG methodology, allowing for items to be generated on-the-fly at the moment of item presentation. The item type was specifically designed such that the structural elements of the item model could be manipulated by a computer algorithm to guide item construction. Using the automatic min-max approach (Arendasy & Sommer, 2010, 2012) as a guiding force, the latent trait, item format, cognitive model, and radicals and formal constraints were specified and deduced by the researcher. Thus, the current research builds upon previous research by creating a unique measure of *Gf* that combines two highly *g*-saturated measures: number series and analogical reasoning tasks. The benefit of using such methodology is that construct relatedness of the measure is built directly into the items generated through the systematic manipulation of the item characteristics thought to relate to item difficulty. Likewise, the elements that could potentially interfere with the cognitive processes involved with solving the items were constrained or omitted. Consequently, the approach taken in the current study allows for the generation of potentially thousands of unique items generated on-the-fly at the moment of presentation, without the need for human review before their administration. The result of this process was a brief 22-item experimental AIG measure of *Gf*, combining two highly *g*-saturated tasks.

The current research was designed to investigate the efficacy of the experimental AIG measure in a sample of adults residing in the United States. In a series of studies, the construct representation, temporal stability, and criterion-relatedness of the experimental AIG measure were examined.

Study 1

In Study 1, along with conventional psychometric analyses, the unidimensionality and construct representation of the experimental AIG measure were assessed. Conventional psychometric statistics indicate that the experimental AIG measure is internally consistent with acceptable discrimination and difficulty values. Likewise, the results of the Martin-Löf and Andersen likelihood ratio tests indicate that the experimental measure data conform well to the Rasch model, supporting its unidimensionality. Using an LLTM analysis (Fischer, 1973; Van den Noortgate, de Boeck, & Meulders, 2003) to test the efficacy of four hypothesized radicals (Consistent Mathematical Operations, Mixed Mathematical Operations, Duplicate Numbers, Flipped Relationships), the results indicated that each significantly contributed to scale difficulty. Therefore, the results of this analysis can be seen as a validation of the use of the hypothesized radicals, thus supporting the construct representation of the experimental AIG measure (Embretson, 1983). Likewise, the results of the LLTM analysis support the inclusion of the radicals not only in the current cognitive model, but also in the generation of future AIG items as they are now calibrated. Furthermore, the hypothesized cognitive model accounted for a substantial portion of the empirically derived difficulty parameter produced by the Rasch model. As such, the proposed model displays adequate content coverage as accounted for by the item radicals. However, an examination of the plot of empirically and LLTM derived difficulty parameters does indicate that the experimental AIG measure tests a limited range of theta with items confined to the range of +2 to -2. As such, the inclusion of a more diverse set of item radicals into the construction of the items may tap a wider breadth of intellectual functioning. While the elements that may improve the content coverage of *Gf* using the present item type are addressed in the Limitations and Future Directions section, the results obtained in Study 1 provide initial evidence for the construct representation of the items generated by the experimental AIG measure.

Study 2

In Study 2 the temporal stability of the experimental AIG measure was assessed by administering the measure on two different occasions approximately one week apart. The results of this analysis indicate that the scores produced by the experimental AIG measure are consistent across testing situations. Specifically, the experimental AIG measure that was administered to participants on two different occasions consisted of a diverse set of items that had varying surface features and resulted in scores that were consistent.

Study 3

In Study 3, the criterion relationships shared between the experimental AIG measure and other established measures of *Gf* were examined. The results of the study indicate that large correlations coefficients were observed between the experimental measure and criterion measures. Likewise, using the total sample of participants, a negative relationship was seen between the experimental AIG measure and age, a phenomenon that has been noted in other investigations of the nature of *Gf* (Cattell, 1943). Taken together, the scores obtained from the experimental AIG measure conform to the scores obtained from other criterion measures of *Gf*, indicating that the scale is tapping aspects of the construct of interest.

Limitations and Future Directions

As in all empirical studies, certain inherent limitations are evident that should be addressed, but also pave the way for future research. Given the many choices that were made in the creation of the experimental AIG measures using an innovative item type, several aspects of item development can be clarified through additional studies to improve the measurement precision of the instrument. For instance, in the current experimental AIG measure, constraints were placed on the size of the randomly generated number used to seed the base number in each term, disallowing numbers to obtain values below zero. Likewise, the same constraint disallowed the numbers that comprised the terms to obtain values greater than 30. Similarly, constraints were placed on the change number that was used to advance each number of the sequence to values of 1 to 4. Future research could relinquish such constraints and then compare the restricted and unrestricted models. As such, additional research is required to assess the need for and effectiveness of limiting seed and change values to such a limited range.

Likewise, while the LLTM analysis described in Study 1 demonstrated that the proposed cognitive model showed substantial content coverage, a more diverse array of potential radical elements is possible. For instance, the current version of the experimental AIG measure limited the size of each analogical reasoning term to three numbers (e.g., 3 – 5 – 7). While this choice was made in the development phase to limit the cognitive complexity of the items generated, as noted by Primi (2001), increasing the number of elements to which examinees must attend in *Gf* tasks is expected to influence item complexity. Therefore, future research on the experimental AIG measure may choose to include number series terms with length as few as two numbers, or increase series length to include four or more numbers. Such changes to the scale may result in a more diverse array of psychometric item attributes, exhibiting a greater range of difficulties.

Additionally, the number of answer choices in the current version of the experimental AIG measure was limited to three. Similar to the length of the number series in each term, varying the number of answer choice from which the participant must choose could lower or raise the cognitive complexity of the items generated as the number of elements from which the examinee must attend changes. Taken together, these two modifications to the experimental AIG measure could serve as radicals in future research, allowing for the production of an extensive array of items exhibiting diverse psychometric properties while still conforming to the *Gf* construct.

The experimental AIG measure also employed a relatively simplistic dichotomous scoring protocol in order to utilize the LLTM analysis. However, future iterations of the experimental AIG measure, or similar measures, could utilize a partial-credit scoring model, allowing for a more diverse range of scores. That is, a more complex scoring algorithm may be applied to the data, allowing item scores to take on a range of values depending on how quickly a correct answer is obtained. As such, participants could be awarded partial credit for answers, better allowing their score to reflect both their speed and accuracy. Hypothetically, a modified scoring scheme such as this may improve the internal consistency, temporal stability, and criterion relatedness of the experimental AIG measure. Simply stated, a modified scoring algorithm may improve the measurement precision of *Gf*.

Importantly, although previous research has shown the value of the data collected from mTurk samples (Buhrmester et al., 2011; Casler et al., 2013), and such crowd-sourcing methods provide psychological researchers an expedient means to obtain variance on a range of psychological attributes (Chandler et al., 2014; Krupnikov & Levine, 2014), their use should be further scrutinized. Logically, the compensation of participants from such sources is tied to how quickly they are able to complete as many of the competing tasks (e.g. the current experimental AIG measure) as are available at the time. Moreover, given the rising costs associated with acquiring participants from such sources (e.g., Bensinger, 2015), the data obtained from these participant pools deserves additional critical analysis as well as potential screening methods to identify high-quality workers.

Likewise, in addition to both video and written instructions detailing the tasks involved in answering the experimental AIG measure, participants were provided an opportunity to practice an unlimited number of items before beginning the actual measure. In addition to becoming familiar with the tasks involved in answering a given item type, practice may allow for a more accurate assessment of an individual's true performance on a given task. As such, given that research has consistently found racial gaps in the scores obtained on cognitive measures (e.g., Roth et al., 2001), limiting their use in organizational settings, the availability of practice items may serve to lessen such gaps. Likewise, such practice items may also serve to reduce examinee apprehension regarding the testing situation and bolster perceptions of fairness. Future research may gauge the impact of practice on *Gf* scores and examinee perceptions, potentially allowing for broader use in selection contexts.

However, given the results, and pending replication, it is possible that an on-the-fly CAT measure can be developed based on the calibrated item radicals of the experimental AIG measure. Using combinations of item radicals, the length of the experimental measure may be greatly reduced, providing a more expedient estimation of *Gf*. Thus, the creation of such an adaptive measure would serve to reduce examinee fatigue while addressing some of the test security threats associated with assessments derived from conventional methods. Likewise, AIG methodology as seen in the current on-the-fly measure also provides stable and effective alternate test forms for use in repeated measures studies and evaluations. As such, researchers and practitioners alike may use these types of scales to evaluate the impact of a variety of psychological interventions. In addition to the test construction and test security issues associated with traditional item construction, researchers may also use these types of measures to assess performance without concerns of item memorization.

Conclusion

The field of cognitive abilities research can be seen as an evolving science. From the early days of Galtonian measures to the advances brought by computerized technology, the field of psychometrics has embraced methodological and technical advances. The advent of AIG methodology serves as the next step in attempting to provide a more complete coverage of the construct space. The current collection of studies introduces an experimental AIG measure as a means to overcome the limitations associated with conventional item creation methods and threats to test security. In sum, the results of these studies highlight the benefits of using AIG methodology to quickly, economically, and effectively generate high-quality on-the-fly *Gf* test items. The experimental AIG measure fulfills the goal of delivering a measure of cognitive ability that is well suited for large-scale cognitive ability assessment via online administration. Thus, as additional research is conducted in the development and calibration of such instruments, researchers and organizations alike may realize the benefit of using AIG methodology to produce effective measures.

REFERENCES

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin, 131*(1), 30-60. doi: 10.1037/0033-2909.131.1.30

Ackerman, P. L., & Heggestad, E. D. (1997). Intelligence, personality, and interests: evidence for overlapping traits. *Psychological Bulletin, 121*(2), 219-245. doi: 10.1037/0033-2909.121.2.219

Al-Saleem, S. M., & Ullah, H. (2014). Security considerations and recommendations in computer-based testing. *The Scientific World Journal, 2014*, 1-7. doi: 10.1155/2014/562787

Alves, C. B., Gierl, M. J., & Lai, H. (2010). *Using automated item generation to promote principled test design and development.* American Educational Research Association, Denver, CO, USA.

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

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrica, 8*, 123–140. doi: 10.1007/BF02291180

Andrews, G., Peter, L., & Teesson, M. (1994). *The measurement of consumer outcomes in mental health: a report to the National Mental Health Information Strategy Committee.* Canberra, Australian: Australian Government Publishing Services.

Arendasy, M. (2000). *Psychometrischer Vergleich computergestützer Vorgabeformen bei Raumvorstellungsaufgaben: Stereoskopischdreidimensionale und herkömmlich-zweidimensionale Darbietung* [Psychometric comparison of computer based presentation modes of spatial ability tasks: Stereoscopic-3D and traditional-2D presentation modes]. PhD Dissertation: University of Vienna.

Arendasy, M. (2005). Automatic generation of Rasch-calibrated items: Figural matrices test GEOM and endless loops Test Ec. *International Journal of Testing, 5*, 197–224. doi: 10.1207/s15327574ijt0503_2

Arendasy, M. E., Hergovich, A., & Sommer, M. (2008). Investigating the 'g'-saturation of various stratum-two factors using automatic item generation. *Intelligence, 36*(6), 574-583. doi: 10.1016/j.intell.2007.11.005

Arendasy, M., Hornke, L. F., Sommer, M., Häusler, J., Wagner-Menghin, M., Gittler, G., … (2007). *Manual Intelligence-Structure-Battery* (INSBAT) Mödling: SCHUHFRIED GmbH.

Arendasy, M., & Sommer, M. (2005). The effect of different types of perceptual manipulations on the dimensionality of automatically generated figural matrices. *Intelligence, 33*, 307–324. doi: 10.1016/j.intell.2005.02.002

Arendasy, M., & Sommer, M. (2007). Automatic generation of quantitative reasoning items: A schema-based isomorphic approach. *Learning and Individual Differences, 17*, 366–383. doi: 10.1027/1614-0001.27.1.2

Arendasy, M., & Sommer, M. (2010). Evaluating the contribution of different item features to the effect size of the gender difference in three-dimensional mental rotation using automatic item generation. *Intelligence, 38*, 574–581. doi: 10.1016/j.intell.2010.06.004

Arendasy, M., & Sommer, M. (2012). Using automatic item generation to meet the increasing item demands of high-stakes assessment. *Learning and Individual Differences, 22*, 112–117. doi: 10.1016/j.lindif.2011.11.005

Arendasy, M. E., & Sommer, M. (2013). Quantitative differences in retest effects across different methods used to construct alternate test forms. *Intelligence, 41*(3), 181-192. doi: 10.1016/j.intell.2013.02.004

Arendasy, M. E., Sommer, M., & Mayr, F. (2012). Using automatic item generation to simultaneously construct German and English versions of a word fluency test. *Journal of Cross-Cultural Psychology, 43*, 464–479. doi: 10.1177/0022022110397360

Arvey, R. D., & Faley, R. H. (1988). *Fairness in selecting employees* (2nd ed.). New York: Addison-Wesley Publishing Company.

Babcock, B., & Weiss, D. J. (2012). Termination criteria in computerized adaptive tests: Do variable-length CATs provide efficient and effective measurement? *Journal of Computerized Adaptive Testing, 1*, 1-18. doi: 10.7333/1212-0101001

Baghaei, P., & Kubinger, K. D. (2015). Linear logistic test modeling with R. *Practical Assessment, Research & Evaluation, 20*(1), 1-11.

Bailey, K. D. (1994). *Typologies and taxonomies: An introduction to classification techniques*. Thousand Oaks, CA: Sage.

Barak, A., & English, N. (2002) Prospects and Limitations of Psychological Testing on the Internet. *Journal of Technology in Human Services, 19*, 65-89, doi: 10.1300/J017v19n02_06

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology, 48*(1), 5-37. doi: 10.1016/j.jsp.2009.10.001

Bartram, D. (2006). *Computer-based testing and the Internet. Issues and advances.* Chichester, UK: John Wiley & Sons.

Baylari, A., & Montazer, G. A. (2009). Design a personalized e-learning system based on item response theory and artificial neural network approach. *Expert Systems with Applications, 36*(4), 8013-8021. doi: 10.1016/j.eswa.2008.10.080

Bejar, I. I. (1990). A generative analysis of a three-dimensional spatial task. *Applied Psychological Measurement, 14*, 237-245. doi: 10.1177/014662169001400302

Bejar, I. I. (2002). Generative testing: From conception to implementation. In S.H. Irvine, & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 199–218). New Jersey: Lawrence Erlbaum Associates.

Bejar, I. I., Lawless, R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *Journal of Technology, Learning, and Assessment, 2*(3). doi: 10.1002/j.2333-8504.2002.tb01890.x

Bell, N. L., Matthews, T. D., Lassiter, K. S., & Leverett, J. P. (2002). Validity of the Wonderlic Personnel Test as a measure of fluid or crystallized intelligence: Implications for career assessment. *North American Journal of Psychology, 4*(1), 113-120.

Bensinger, G. (2015). *Amazon's Mechanical Turk fee hike irks researchers*. Retrieved from http://blogs.wsj.com/digits/2015/06/23/amazons-mechanical-turk-fee-hike-irks-researchers/

Bertua, C., Anderson, N., & Salgado, J. (2005). The predictive validity of cognitive ability tests. A UK meta-analysis. *Journal of Occupational and Organisational Psychology, 78*, 387–409. doi: 10.1348/096317905X26994

Bickley, P. G., Keith, T. Z., & Wolfle, L. M. (1995). The three-stratum theory of cognitive abilities: Test of the structure of intelligence across the life span. *Intelligence, 20*(3), 309-328. doi: 10.1016/0160-2896(95)90013-6

Blanchette, I., & Dunbar, K. (2002). Representational change and analogy: how analogical inferences alter target representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(4), 672-685. doi: 10.1037//0278-7393.28.4.672

Boake, C. (2002). From the Binet-Simon to the Wechsler-Bellevue: Tracing the history of intelligence testing. *Journal of Clinical and Experimental Neuropsychology, 24*(3), 383-405.

Boring, E. G. (1923). Intelligence as the tests test it. *New Republic, 36*, 35–37. doi: 10.1037/11491-017

Borman, W., Hanson, M., Oppler, S., Pulakis, E., & White, L. (1993). Role of early supervisory experience in supervisor performance. *Journal of Applied Psychology, 78*, 443–449. doi: 10.1037/0021-9010.78.3.443

Bors, D. A., & Forrin, B. (1995). Age, speed of information processing, recall, and fluid intelligence. *Intelligence, 20*(3), 229-248.

Bradley, S. (2010). *The meaning of shapes: Developing visual grammar.* Vanseo Design. Retrieved from http://vanseodesign.com/web-design/visual-grammar-shapes/

Breithaupt, K., Ariel, A. A., & Hare, D. R. (2009). *Assembling an inventory of multistage adaptive testing systems. In Elements of adaptive testing* (pp. 247-266). Springer New York.

Brody, N. (1992). *Intelligence* (2nd ed.). Academic Press, New York.

Brody, N. (2003). Construct validation of the Sternberg Triarchic abilities test: Comment and reanalysis. *Intelligence, 31*(4), 319-329. doi: 10.1016/S0160-2896(01)00087-3

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*(1), 3-5. doi: 10.1177/1745691610393980

Burlingame, G. M., Lambert, M. J., Reisinger, C. W., Neff, W. M., & Mosier, J. (1995). Pragmatics of tracking mental health outcomes in a managed care setting. *The Journal of Mental Health Administration, 22*(3), 226-236. doi: 10.1007/BF02521118

Burt, C. (1949). The structure of the mind: A review of the results of factor analysis. *British Journal of Educational Psychology*, *19*, 176-199. doi: 10.1111/j.2044-8279.1949.tb01621.x

Bynum, W. F. (2002). The childless father of eugenics. *Science, 296*, 472. doi: 10.1126/science.1069041

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105. doi: 10.1037/h0046016

Campbell, J. P. (1996). Group differences and personnel decisions: Validity, fairness and affirmative action. *Journal of Vocational Behaviour, 49*, 122-158. doi: 10.1006/jvbe.1996.0038

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices test. *Psychological Review, 97*(3), 404–431. doi: 10.1037/0033-295X.97.3.404

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

Carroll, J. B. (1997). Psychometrics, intelligence, and public perception. *Intelligence, 24*, 25–52. doi: 10.1016/S0160-2896(97)90012-X

Carroll, J. B. (2003). The higher-stratum structure of cognitive abilities: Current evidence supports g and about ten broad factors. In H. Nyborg (Ed.), *The scientific study of general intelligence: Tribute to Arthur R. Jensen* (pp. 5-21). New York: Pergamon.

Carroll, J. B. (2005). The three-stratum theory of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Eds.), Cont*emporary intellectual assessment: Theories, tests and issues* (2nd ed., pp. 69–76). New York: Guilford.

Cascio, W. F., & Aguinis, H. (2005). *Applied Psychology in Human Resource Management.* New Jersey: Prentice Hall.

Casey, M. B., Nuttall, R., Pezaris, E. and Benbow, C. (1995). The influence of spatial ability of gender differences in mathematics college entrance test scores across diverse samples. *Developmental Psychology, 31*(4), 697-705. doi: 10.1037/0012-1649.31.4.697

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior, 29*(6), 2156-2160.

Castella-Roca, J., Herrera-Joancomarti, J., & Dorca-Josa, A. (2006, April). A secure e-exam management system. In *First International Conference on Availability, Reliability and Security* (ARES'06). IEEE.

Cattell, J. M. (1890). Mental tests and measurements. *Mind, 15*, 373-381. doi: 10.2307/1411718

Cattell, J. M. (1928). Early psychological laboratories. *Science, 67*, 543-548. doi: 10.1126/science.67.1744.543

Cattell, R. B. (1940). A culture-free intelligence test. *Journal of Educational Psychology, 31*, 176–199. doi: 10.1037/h0059043

Cattell, R. B. (1941). Some theoretical issues in adult intelligence testing. *Psychological Bulletin, 38*, 592.

Cattell, R. B. (1943). The measurement of adult intelligence. *Psychological Bulletin, 40*(3), 153-193. doi:10.1037/h0059973

Cattell, R. B. (1957). *Personality and motivation structure and measurement*. New York, NY: World Book.

Cattell, R. B. (1971). *Abilities: Their structure, growth, and action.* Boston: Houghton Mifflin.

Cattell, R. B., & Cattell, A. K. S. (1963). *Culture Fair Intelligence Test.* Champaign, IL: Institute for Personality and Ability Testing.

Ceci, S. J. (1991). How much does schooling influence general intelligence and its cognitive components? A reassessment of the evidence. *Developmental Psychology, 27*, 703-722. doi: 10.1037/0012-1649.27.5.703

Chamorro-Premuzic, T., & Furnham, A. (2010). *The psychology of personnel selection.* Cambridge University Press.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*(1), 112–130. doi: 10.3758/s13428-013-0365-7

Chang, S., & Ansley, T. N. (2003). A comparative study of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 40*, 71-103. doi: 10.1111/j.1745-3984.2003.tb01097.x

Chernyshenko, O. S., & Stark, S. (2015). Mobile psychological assessment. In F. Drasgow, Ed., *Technology and Testing: Improving Educational and Psychological Measurement* (Vol. 2). Hoboken, NJ: Wiley-Blackwell.

Clayes, G. (2001). Introducing Francis Galton, 'Kantsaywhere' and 'The Donoghues of Dunno Weir.' *Utopian Studies, 12*(2), 188-190.

Cohen, R. J., & Swerdlik, M. E. (2009). *Psychological testing and assessment: An introduction to tests and measurements* (7th ed.). New York, NY: McGraw-Hill.

Colom, R., & García-López, O. (2002). Sex differences in fluid intelligence among high school graduates. *Personality and Individual Differences*, *32*(3), 445-451. doi: 10.1016/S0191-8869(01)00040-X

Condon, D. M., & Revelle, W. (2014). The International Cognitive Ability Resource: Development and initial validation of a public-domain measure. *Intelligence, 43*, 52-64. doi: 10.1016/j.intell.2014.01.004

Cook, L. L., & Eignor, D. R. (1991). NCME Instructional module: IRT equating methods. *Educational Measurement: Issues and Practice, 10*, 37–45.

Costa, P. T. Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual.* Odessa, Florida: Psychological Assessment Resources, Inc.

Croft, M. (2014). The end of erasures: Updating test security laws and policies for computerized testing. *ACT Research & Policy*, 1-5. Retrieved from http://www.act.org/research/policymakers/pdf/EndofErasures.pdf

Cronbach, L. J. (1990). *Essentials of psychological testing.* (5th ed.), New York: Harper & Row.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302. doi: 10.1037/h0040957

Dahlstrom, W. G. (1985). The development of psychological testing. In G. A. Kimble and K. Schlesinger (Eds.), *Topics in the history of psychology*, 2, 63-114.

Darwin, C. (1859). *On the origin of species by means of natural selection.* London: John Murry.

Deary, I. (2004). *Looking down on intelligence*. Oxford, UK: Oxford University Press.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach.* New York: Springer.

DeVellis, R. F. (2012). *Scale development: Theory and applications* (3rd ed.). Los Angeles, CA: Sage.

DiCerbo, K. E., & Behrens, J. T. (2012). Implications of the digital ocean on current and future assessment. In R. Lissitz & H. Jiao (Eds.), *Computers and their impact on state assessment: Recent history and predictions for the future* (pp. 273–306). Charlotte, NC: Information Age Publishing.

Doebler, A., & Holling, H. (2015). A processing speed test based on rule-based item generation: An analysis with the Rasch Poisson Counts model. *Learning and Individual Differences*, 1-8. doi: 10.1016/j.lindif.2015.01.013

Doebler, P. (2015). *The figural analogies development process. ICAR News, 1*, 4-7.

Downing, S. M. (2003). Item response theory: applications of modern test theory for assessments in medical education. *Medical Education, 37*, 739-745. doi: 10.1046/j.1365-2923.2003.01587.x

Downing, S. M., & Haladyna, T. M. (2006). *Handbook of test development.* Lawrence Erlbaum Associates Publishers.

Downing, S. M., & Yudkowsky, R. (2009). *Assessment in health professions education.* London: Routledge.

Drasgow, F., Luecht, R. M., & Bennett, R. (2006). Technology and testing. In: Brennan RL, ed. *Educational Measurement* (4th ed., pp. 471-516). Washington, DC: American Council on Education 2006.

Drasgow, F., & Mattern, K. (2006). New tests and new items: Opportunities and issues. In D. Bartram & R.K. Hambleton (Eds.), *Computer-based testing and the Internet: Issues and advances* (pp. 59-75). Chichester: Wiley.

Drasgow, F., Nye, C. D., Guo, J., & Tay, L. (2009). Cheating on proctored tests: The other side of the unproctored debate. *Industrial and Organizational Psychology, 2*(1), 46-48. doi: 10.1111/j.1754-9434.2008.01106.x

Drasgow, F., & Olson-Buchanan, J. B. (Eds.). (1999). *Innovations in computerized assessment*. Psychology Press.

DuBois, P. H. (1970). *A history ofpsychological testing*. Boston: Allyn & Bacon.

Duncan, J., Seitz, R. J., Kolodny, J., Bor, D., Herzog, H., Ahmed, A.,...Emslie, H. (2000). *A neural basis for general intelligence*. Science, *289*, 457-460. doi: 10.1016/S0002-9394(00)00752-2

Duncan, O. D. (1984). *Notes on social measurement: Historical and critical*. New York: Russell Sage.

Duran, R., Powers, D., & Swinton, S. (1987). Construct validity of the GRE analytical test: A resource document. *ETS Research Report Series, 1987*(1), i-91.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Manual for kit of factor-referenced cognitive tests*. Princeton, NJ: Educational testing service.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179–197.

Embretson, S. E. (2002). Generating abstract reasoning items with cognitive theory. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 219-250). Mahwah, NJ: Erlbaum.

Embretson, S. E. (2005). Measuring human intelligence with artificial intelligence. In R. J. Sternberg, & J. E. Pretz (Eds.), *Cognition and intelligence* (pp. 251–267). New York: Cambridge University Press.

Embretson, S. E., & Daniel, R. C. (2008). Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychological Science Quarterly, 50*, 328-344.

Evans, T. G. (1968). Program for the solution of a class of geometric-analogy intelligent-test questions. In M. Minsky (Ed.), *Semantic information processing.* Cambridge, MA: MIT Press.

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

Feldman, J. J. (1966). *The dissemination of health information: A case study in adult learning.* Chicago: Aldine.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: Sage.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 36*, 207–220. doi: 10.1016/0001-6918(73)90003-6

Fischer, G. H. (1995). The linear logistic test model. In G.H. Fischer, & I. W. Molenaar (Eds.), *Rasch models. Foundations, recent developments, and applications* (pp. 157–180). New York: Springer.

Fischer, G. H. (2005). Linear logistic test models. In *Encyclopedia of Social Measurement*, *2*, 505-514. doi: 10.1016/B0-12-369398-5/00453-9

Flanagan, D. P. (2000).Wechsler-based CHC cross-battery assessment and reading achievement: Strengthening the validity of interpretations drawn from Wechsler test scores. *School Psychology Quarterly, 15*(3), 295–329. doi: 10.1037/h0088789

Forrest, D. W. (1974). *Francis Galton: The Life and Work of a Victorian Genius.* New York: Paul Elek Ltd.

Foster, D. F. (2010). Worldwide testing and test security issues: Ethical challenges and solutions. *Ethics & behavior, 20*(3-4), 207-228. doi: 10.1080/10508421003798943

Freedheim, D. K., & Weiner, I. B. (2003). *Handbook of Psychology, History of Psychology* (Vol. 1). John Wiley & Sons.

French, R. M. (2002). The computational modeling of analogy-making. *Trends in cognitive Sciences, 6*(5), 200-205. doi: 10.1016/S1364-6613(02)01882-X

Freund, P. A., Hofer, S., & Holling, H. (2008). Explaining and controlling for the psychometric properties of computer-generated figural matrix items. *Applied Psychological Measurement, 32*(3), 195-210. doi: 10.1177/0146621607306972

Freund, P. A., & Holling, H. (2011). How to get really smart: Modeling retest and training effects in ability testing using computer-generated figural matrix items. *Intelligence, 39*(4), 233-243. doi: 10.1016/j.intell.2011.02.009

Furnham, A. (2008). *Personality and intelligence at work: Exploring and explaining individual differences at work.* London: Routledge.

Futschek, K. (2014). Actual type-I-and type-II-risk of four different model tests of the Rasch model. *Psychological Test and Assessment Modeling, 56*(2), 168-177.

Galton, F. (1883). *Inquiries into human faculty and its development.* London: Macmillan.

Gardner, H. (1999). *Intelligence reframed: Multiple intelligence for the 21st century.* New York: Basic Books.

Gardner, H. (2011). *Frames of mind: The theory of multiple intelligences.* New York, NY: Basic Books.

Gardner, H., & Hatch, T. (1989). Educational implications of the theory of multiple intelligences. *Educational Researcher, 18*(8), 4-10. doi: 10.3102/0013189X018008004

Geerlings, H., Glas, C. A. W., & van der Linden, W. J. (2011). Modeling rule-based item generation. *Psychometrika, 76*, 337-359. doi: 10.1007/S11336-011-9204-X

Geerlings, H., van der Linden, W. J., & Glas, C. A. (2012). Optimal test design with rule-based item generation. *Applied Psychological Measurement, 37*(2), 140-161. doi: 10.1177/0146621612468313

Ghiselli, E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences.* San Francisco, CA: Freeman

Gierl, M. J., Ball, M. M., Vele, V., & Lai, H. (2015). A Method for Generating Nonverbal Reasoning Items Using n-Layer Modeling. In *Computer Assisted Assessment. Research into E-Assessment* (pp. 12-21). Springer International Publishing. doi: 10.1007/978-3-319-27704-2_2

Gierl, M. J., & Haladyna, T. M. (Eds.). (2012). *Automatic item generation: Theory and practice.* London: Routledge.

Gierl, M. J., & Lai, H. (2012). The role of item models in automatic item generation. *International Journal of Testing, 12*(3), 273-298. doi: 10.1080/15305058.2011.635830

Gierl, M. J., Lai, H., & Turner, S. (2012). Using automatic item generation to create multiple-choice items for assessments in medical education. *Medical Education, 46*, 757-765. doi: 10.1111/j.1365-2923.2012.042989.x

Gittler, G. (1990). *Dreidimensionaler Würfeltest—Ein Rasch-skalierter Test zur Messung des räumlichen Vorstellungsvermögens. Theoretische Grundlagen und Manual* [Three dimensional cubes test—A Rasch-calibrated test for the measurement of spatial ability. Theoretical background and Manual] Weinheim: Beltz.

Gittler, G., & Arendasy, M. (2003). Endlosschleifen: Psychometrische Grundlagen des Aufgabentyps EP [Endless Loops: Psychometric foundations of the item type EP]. *Diagnostica, 49*, 164–175.

Glas, C. A., & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement, 27*, 247–261. doi: 10.1177/0146621603027004001

Goldman, S. R., & Pellegrino, J. W. (1984). Deductions about induction: Analyses of developmental and individual differences. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 2, pp. 149–197). Hillsdale, NJ: Erlbaum.

Gordon, R. A., Lewis, M. A., & Quigley, A. M. (1988). Can we count on muddling through the g crisis in employment? *Journal of Vocational Behavior, 33*(3), 424-451. doi: 10.1016/0001-8791(88)90049-8

Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement, 30*, 394-411. doi: 10.1177/0146621606288554

Gottfredson, L. S. (1997). Why g matters: The complexity of everyday life. *Intelligence, 24*(1), 79-132. doi: 10.1016/S0160-2896(97)90014-3

Gottfredson, L. S. (1998, Winter). The general intelligence factor. *Scientific American Presents, 9*(4), 24-29.

Gottfredson, L. S. (2002). Where and why g matters: Not a mystery. *Human Performance, 15*, 25–46. doi: 10.1207/S15327043HUP1501&02_03

Gottfredson, L. S. (2003). g, jobs, and life. In H. Nyborg (Ed.), *The scientific study of general intelligence: Tribute to Arthur R. Jensen* (pp. 293-342). New York: Pergamon.

Gottfredson, L. S. (2004). Intelligence: Is it the epidemiologists' elusive "fundamental cause" of social class inequalities in health? *Journal of Personality and Social Psychology, 86*(1), 174-199. doi: 10.1037/0022-3514.86.1.174

Grabner, R. H., Fink, A., Stipacek, A., Neuper, C., & Neubauer, A. C. (2004). Intelligence and working memory systems: evidence of neural efficiency in alpha band ERD. *Cognitive Brain Research, 20*(2), 212-225. doi: 10.1016/j.cogbrainres.2004.02.010

Grover, S. (1996). The business necessity defense in disparate impact discrimination cases. *Georgia Law Review, 30*, 387-429.

Guilford, J. P. (1954). Psychometric methods. New York, NY: McGraw-Hill.

Guilford, J. P. (1967). *The Nature of Human Intelligence*. New York: McGraw-Hill.

Guilford, J. P. (1988). Some changes in the structure-of-intellect model. Educational and *Psychological Measurement, 48*, 1-4. doi:10.1177/ 001316448804800102

Guion, R. M., & Highouse, S. (2006). *Essentials of personnel assessment and selection.* Mahwah, NJ: Erlbaum.

Guo, J., Tay, L., & Drasgow, F. (2009). Conspiracies and test compromise: An evaluation of the resistance of test systems to small-scale cheating. *International Journal of Testing, 9*(4), 283-309.

Gustafsson, J. E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence, 8*, 179–203. doi: 10.1016/0160-2896(84)90008-4

Gustafsson, J. E. (1989). Broad and narrow abilities in research on learning and instruction. In: R. Kanfer, P. L. Ackerman, & R. Cudeck (Eds.), *Abilities, motivation, and methodology: The Minnesota Symposium on Learning and Individual Differences* (pp. 203–237). Hillsdale, NJ: Erlbaum.

Gustafsson, J. E. (2001). On the hierarchical structure of ability and personality. In: J. M. Collis & S. Messick (Eds), *Intelligence and personality: Bridging the gap in theory and measurement* (pp. 25–42). Mahwah, NJ: Erlbaum.

Halpern, D. F. (1997). Sex differences in intelligence: Implications for education. *American Psychologist, 52*(10), 1091-1102. doi: 10.1037/0003-066X.52.10.1091

Harrell, T. W., & Harrell, M. S. (1945). Army General Classification Test scores for civilian occupations. *Educational and Psychological Measurement, 5*, 229–239. doi: 10.1177/001316444500500303

Hartigan, J. A., & Wigdor, A. K. (1989). *Fairness in employment testing: Validity generalization, minority issues and the General Aptitude Test Battery.* Washington, DC: National Academy Press.

Hausdorf, P. A., LeBlanc, M. M., & Chawla, A. (2003). Cognitive ability testing and employment selection: Does test content relate to adverse impact? *Applied H.R.M. Research, 7*(1-2), 41-48.

Henryssen, S. (1971). Gathering, analyzing, and using data on test items. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). Washington, DC: American Council on Education.

Hergenhahn, B. R. (2009). *An introduction to the history of psychology* (6th ed.). Belmont, CA: Wadsworth Cengage Learning.

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

Hicks, K. L., Harrison, T. L., & Engle, R. W. (2015). Wonderlic, working memory capacity, and fluid intelligence. *Intelligence, 50*, 186-195. doi: 10.1016/j.intell.2015.03.005

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods, 1*(1), 104-121. doi: 10.1177/109442819800100106

Holling, H., Bertling, J. P., & Zeuch, N. (2009). Automatic item generation of probability word problems. *Studies in Educational Evaluation, 35*(2-3), 71–76. doi: 10.1016/j.stueduc.2009.10.004

Holyoak, K. J. (2005). Analogy. In K. J. Holyoak & R. G. Morrison (Eds.) *The Cambridge handbook of thinking and reasoning* (pp. 117-142). Cambridge University Press.

Holyoak, K. J., & Morrison, R. G. (2005). *The Cambridge handbook of thinking and reasoning*. Cambridge University Press.

Hopkins, W. G. (2002). A scale of magnitudes for effect statistics. In *A New View of Statistics*. Retrieved from http://www.sportsci.org/resource/stats/effectmag.html

Horn, J. L. (1968). Organization of abilities and the development of intelligence. *Psychological Review, 75*, 242–259. doi: 10.1037/h0025662

Horn, J. L. (1988). Thinking about human abilities. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate psychology* (Rev. ed., pp. 645–685). New York, NY: Academic Press.

Horn, J. L. (1991). Measurement of intellectual capabilities: A review of theory. In K. S. McGrew, J. K. Werder, & R. W. Woodcock, *Woodcock–Johnson technical manual* (pp. 197–232). Itasca, IL: Riverside.

Horn, J. L., & Blankson, N. (2005) Foundations for better understanding of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 41–68). New York, NY: Guilford Press.

Horn, J. L., & McArdle, J. J. (2007). Understanding human intelligence since Spearman. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 205-247). Mahwah, NJ: Erlbaum.

Horn, J. L., & Noll, J. (1997). Human cognitive capabilities: Gf-Gc theory. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), Co*ntemporary intellectual assessment: Theories, tests, and issues* (pp. 53-91). New York: Guilford.

Hornke, L. F., & Habon, M. W. (1986). Rule-based item bank construction and evaluation within the linear logistic framework. *Applied Psychological Measurement, 10*, 369–380. doi: 10.1177/014662168601000405

Hothersall, D. (1995). *History of Psychology*. New York: McGraw-Hill, Inc.

Howell, D. C. (2013). *Statistical methods for psychology* (8th Ed.). Belmont, CA: Wadsworth.

Hulsheger, U., Maier, G., & Stumpp, T. (2007). Validity of general mental ability for the prediction of job performance and training success in Germany. *International Journal of Selection and Assessment, 15*, 3–18. doi: 10.1111/j.1468-2389.2007.00363.x

Humphreys, L. G. (1962). The organization of human abilities. *American Psychologist, 17*, 475–483. doi: 10.1037/h0041550

Hunt, E. (2001). Multiple views of multiple intelligence. [Review of Intelligence reframed: Multiple intelligence in the 21st century.] *Contemporary Psychology, 46*, 5–7. doi: 10.1037/002513

Hunter, J. E. (1980). *Test validation for 12,000 jobs: An application of synthetic validity and validity generalization to the General Aptitude Test Battery (GATB).* Washington, DC: U. S. Employment Service.

Hunter, J. E. (1989). *The Wonderlic Personnel Test as a predictor of training success and job performance*. Libertyville, IL: Wonderlic.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72–98. doi: 10.1037/0033-2909.96.1.72

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology, 91*, 594–612. doi: 10.1037/0021-9010.91.3.594

Irvine, P. (1986). Sir Francis Galton (1822-1911). *Journal of Special Education, 20*(1). doi: 10.1177/002246698602000102

Irvine, S. (2002). The foundations of item generation for mass testing. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp.3-32). Hillsdale, NJ: Erlbaum.

Jacobs, L. C., & Chase, C. I. (1992). *Developing and using tests effectively: A guide for faculty.* (pp. 168–177) San Francisco, CA: Jossey-Bass.

Jacobs, P. J., & Vandeventer, M. (1972). Evaluating the teaching of intelligence. *Educational and Psychological Measurement, 32*, 235–248.

Jensen, A. R. (1980). *Bias in mental testing.* New York: Free Press.

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

Jensen, A. R. (2004). Obituary—John Bissell Carroll. *Intelligence, 32*(1), 1–5.

Karim, M. N., Kaminsky, S. E., & Behrend, T. S. (2014). Cheating, reactions, and performance in remotely proctored testing: An exploratory experimental study. *Journal of Business and Psychology, 29*(4), 555-572. doi: 10.1007/s10869-014-9343-z

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

Keith, T. Z., Kranzler, J. H., & Flanagan, D. P. (2001). What does the Cognitive Assessment System (CAS) measure? Joint confirmatory factor analysis of the CAS and the Woodcock–Johnson Tests of Cognitive Ability (3rd ed.). *School Psychology Review, 30*, 89-119.

Keith, T. Z., & Reynolds, M. R. (2010). Cattell–Horn–Carroll abilities and cognitive tests: What we've learned from 20 years of research. *Psychology in the Schools, 47*(7), 635-650. doi: 10.1002/pits.20496

Kline, P. (1999). *The handbook of psychological testing* (2nd ed.). London: Routledge.

Kranzler, J. H., & Keith, T. Z. (1999). Independent confirmatory factor analysis of the Cognitive Assessment System (CAS): What does the CAS measure? *School Psychology Review, 28*, 117-144.

Krupnikov, Y., & Levine, A. S. (2014). Cross-sample comparisons and external validity. *Journal of Experimental Political Science*, *1*, 59–80. doi: 10.1017/xps.2014.7

Kvist, A. V., & Gustafsson, J. E. (2008). The relation between fluid intelligence and the general factor as a function of cultural background: A test of Cattell's investment theory. *Intelligence, 36*(5), 422-436. doi: 10.1016/j.intell.2007.08.004

Kyle, T. (2002). *Cheating scandal rocks GRE, ETS.* Retrieved from http://thedartmouth.com/2002/08/09/cheating-scandal-rocks-gre-ets/

Kyllonen, P. (2002). Item generation for repeated testing of human performance. In S. Irvine & P. Kyllonen (Eds.), *Item generation for test development*. (pp. 251-276). Mahwah, NJ: Lawrence Earlbaum Associates.

Lai, H., Alves, C., & Gierl, M. J. (2009). Using automatic item generation to address item demands for CAT. In *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.*

Lamb, K. (1994). Genetics and Spearman's "g" factor. *Mankind Quarterly, 34*(4), 379–391.

Larson, G. (1994) Armed services vocational aptitude battery. In R.J. Sternberg (Ed.), *Encyclopedia of intelligence* (Vol. 1, pp. 121-124.) New York: Macmillan.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*, 2nd ed. Hoboken, NJ: Wiley.

Lubinski, D. (2004). Introduction to the special section on cognitive abilities. *Journal of Personality and Social Psychology, 86*, 96–111.

Lubinski, D., & Benbow, C. P. (1995). An opportunity for empiricism: Review of Howard Gardner's Multiple intelligences: The theory in practice. *Contemporary Psychology, 40*, 935-938. doi: 10.1037/004016

Luria, A. R. (1966). Human brain and psychological processes. New York, NY: Harper & Row.

Lynn, R., & Irwing, P. (2004). Sex differences on the progressive matrices: A meta-analysis. *Intelligence*, *32*(5), 481-498. doi: 10.1016/j.intell.2004.06.008

Mader, S. S. (1996) *Biology*. (5th ed.), Dubuque, IA: William C Brown.

Mair, P., Hatzinger, R., & Maier, M. J. (2013). *eRm: Extended Rasch Modeling* [Computer software]. R package version 0.15-3.

Martin-Löf, P. (1973). *Statistika modeler: Anteckningar fran seminarier lasa°ret 1969-1970, utarbetade av Rolf Sundberg. Obetydligt a¨ndrat nytryk, Oktober 1973* [Statistical models: Notes from seminars 1969-1970, prepared by Rolf Sundberg]. Stockholm, Sweden: Institutet fo¨ r Fo¨rsa¨kringsmatematik och Matematisk Statistisk vid Stockhokms Universitet.

Matthews, T. D., & Lassiter, K. S. (2007). What does the Wonderlic Personnel Test measure? *Psychological Reports, 100*(3), 707-712. doi: 10.2466/pr0.100.3.707-712

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review, 15*, 28-50. doi: 10.1177/1088868310366253

McDaniel, M. A., & Banks, G. C. (2010). General cognitive ability. In J. Scott and D. Reynolds (Eds.) *The Handbook of Workplace Assessment: Selecting and Developing Organizational Talent*. Hoboken, NJ: Wiley. 61-80.

McGrew, K. S. (1997). Analysis of the major intelligence batteries according to a proposed comprehensive Gf–Gc framework. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 151−179). New York: Guilford.

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), 1-10. doi:10.1016/j.intell.2008.08.004

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

McGuire, F. (1994). Army alpha and beta tests of intelligence. In R. J. Sternberg (Ed.), *Encyclopedia of intelligence* (Vol 1, pp. 125-129.) New York: Macmillan.

Mead, A., & Drasgow, F. (1993). Equivalence of Computerized and paper-pencil cognitive ability tests: A meta- analysis. *Psychological Bulletin, 114*(3). 449-458. doi: 10.1037/0033-2909.114.3.449

Meyer, J. P., & Zhu, S. (2013). Fair and equitable measurement of student learning in MOOCs: An introduction to item response theory, scale linking, and score equating. *Research & Practice in Assessment, 8*, 26-39.

Miller, D. (1996). Configurations revisited. *Strategic Management Journal, 17*(7), 505–512. doi: 10.1002/(SICI)1097-0266(199607)17:7%3C505::AID-SMJ852%3E3.3.CO;2-9

Minton, H. L. (1988) *Lewis M. Terman*. New York, New York: University Press.

Morley, M. E., Bridgeman, B., & Lawless, R. R. (2004). *Transfer between variants of quantitative items*. Princeton, N.J.: Educational Testing Service (GRE Board Research Report No. 00-06R).

Mulholland, T. M., Pellegrino, J. W., & Glaser, R. (1980). Components of geometric analogy solution. *Cognitive Psychology, 12*(2), 252-284. doi: 10.1016/0010-0285(80)90011-0

Naglieri, J. A., & Das, J. P. (1997). *Das-Naglieri cognitive assessment system*. Itasca, IL: Riverside.

Naglieri, J. A., Das, J. P., & Goldstein, S. (2014). *Cognitive assessment system–2.* Austin, TX: ProEd.

Naglieri, J. A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., & Velasquez, R. (2004). Psychological testing on the Internet: New problems, old issues. *American Psychologist. 3*, 150-162. doi: 10.1037/0003-066X.59.3.150

Neisser, U. (1967). *Cognitive psychology.* New York: Appleton Century Crofts.

Neisser, U., Boodoo, G., Bouchard, T., Boykin, A., Brody, N., Ceci, S., Halpern, D., Loehlin, J., Perloff, R., Sternberg, R., & Urbina, S. (1996). Intelligence: knowns and unkowns. *American Psychologist*, *51*(2), 77–101.

Nettlebeck, T., & Wilson, C. (2005). Intelligence and IQ: What teachers should know? *Educational Psychology, 25*, 609–630. doi: 10.1080/01443410500344696

Newman, D. A. (2003). Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. *Organizational Research Methods, 6*(3), 328-362. doi: 10.1177/1094428103254673

Nisbett, R. E., Aronson, J., Blair, C., Dickens, W., Flynn, J., Halpern, D. F., & Turkheimer, E. (2012). Intelligence: New findings and theoretical developments. *American Psychologist, 67*(2), 130-159. doi: 10.1037/a0026699

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

Nunnally, J. S., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

O'Toole, B. I., & Stankov, L. (1992). Ultimate validity of psychological tests. *Personality and Individual Differences, 13*, 699–716.

Ones, D., Viswesvaran, C., & Dilchert, S. (2006). Cognitive ability in selection decisions. In D. Wilhelm & R. Engle (Eds.), *Understanding and measuring intelligence*. London: Sage.

Parshall, C. G., & Harmes, J. C. (2009). Improving the quality of innovative item types: Four tasks for design and development. *Journal of Applied Testing Technology, 10* (1), 1-20.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao, &S. Sinharay (Eds.), *Handbook of statistics. Psychometrics* (Vol. 26, pp. 125–167) North Holland: Elsevier.

Penrose, L. S., & Raven, J. C. (1936). A new series of perceptual tests: Preliminary communication. *British Journal of Medical Psychology, 16*, 97-104. doi: 10.1111/j.2044-8341.1936.tb00690.x

PHP.net (2016). *PHP Manual*. Retrieved from http://php.net/manual/en/index.php

Poinstingl, H. (2009). The Linear Logistic Test Model (LLTM) as the methodological foundation of item generating rules for a new verbal reasoning test. *Psychological Test and Assessment Modeling, 51*(2), 123.

Porter, T. M. (1995). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton, NJ: Princeton University Press.

Prabhakaran, V., Smith, J. A., Desmond, J. E., Glover, G. H., Gabrieli, J. D. (1997). Neural substrates of fluid reasoning: An fMRI study of neocortical activation during performance of the Raven's Progressive Matrices Test. *Cognitive Psychology, 33*(1), 43-63. doi: 10.1006/cogp.1997.0659

Primi, R. (2001). Complexity of geometric inductive reasoning tasks: Contribution to the understanding of fluid intelligence. *Intelligence, 30*(1), 41-70. doi: 10.1016/S0160-2896(01)00067-8

Randall, J., Sireci, S., Li, X., & Kaira, L. (2012). Evaluating the Comparability of Paper- and Computer-Based Science Tests Across Sex and SES Subgroups. *Educational Measurement: Issues and Practice, 31*(4), 2-12.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: The University of Chicago Press.

Raven, J. C. (1938). *Progressive matrices: A perceptual test of intelligence*. London: H. K. Lewis.

Raven, J., & Court, J. H. (1989). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Research supplement no. 4: Additional national and American norms, and summaries of normative, reliability, and validity studies.* Oxford, England: Oxford Psychologists Press/San Antonio, TX: The Psychological Corporation.

Raven, J. C., Court, J. H., & Raven, J. (1978). *Manual for Raven's Progressive Matrices and Vocabulary Scales*. London, UK: H.K. Lewis.

Raven, J. C., Court, J. H., & Raven, J. (1998). Ma*nual for Raven's Progressive Matrices and Vocabulary Scales (Section 3)*. Oxford: Oxford Psychologist Press.

Raven, J., Raven, J. C., & Court, J. H. (2003). *Manual for Raven's Progressive Matrices and Vocabulary Scales*. San Antonio, TX: Harcourt Assessment.

Redick, T. S., Unsworth, N., Kelly, A. J., & Engle, R. W. (2012). Faster, smarter? Working memory capacity and perceptual speed in relation to fluid intelligence. *Journal of Cognitive Psychology, 24*(7), 844-854. doi: 10.1080/20445911.2012.704359

Ree, M. J., & Earles, J. A. (1993). g is to psychology what carbon is to chemistry: A reply to Sternberg and Wagner, McClelland, and Calfee. *Current Directions in Psychological Science, 2*, 11–12. doi: 10.1111/1467-8721.ep10770509

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predictive job performance: Not much more than *g*. *Journal of Applied Psychology, 79*, 518-524.

Reynolds, D. H., & Dickter, D. N. (2010). Technology and employee selection. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of Employee Selection.* Clifton, NJ: Psychological Press.

Robin, N., & Holyoak, K. J. (1995). Relational complexity and the functions of prefrontal cortex. In M. S. Gazzaniga (Ed.). *The cognitive neurosciences* (pp. 987–997). Cambridge, MA: MIT Press.

Roid, G. H. (2005). *Stanford-Binet intelligence scales* (5th ed.). Austin, Tx: Pro-Ed, Inc.

Rost, J. (1982). An Unconditional Likelihood Ratio for Testing Item Homogeneity in the Rasch Model. *Education Research and Perspectives*, *9*(1), 7-17.

Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S., & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology, 54*, 297-330. doi: 10.1111/j.1744-6570.2001.tb00094.x

Rudner, L. M. (2009). Implementing the graduate management admission test computerized adaptive test. In *Elements of adaptive testing* (pp. 151-165). Springer New York.

Rushton, J. P., & Jensen, A. R. (2005). Thirty years of research on race differences in cognitive ability. *Psychology, Public Policy, and Law, 11*, 235–294. doi:10.1037/1076-8971.11.2.235

Sackett, P. R., & Ellingson, J. E. (1997). The effects of forming multi-predictor composites on group differences and adverse impact. *Personnel Psychology, 50,* 707-721. doi: 10.1111/j.1744-6570.1997.tb00711.x

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., & De Fruyt, F. (2003). International validity generalisation of GMA and cognitive abilities. *Personnel Psychology, 56*, 573–605.

Salgado, J. F., Moscoso, S., & Anderson, N. (2016). Corrections for criterion reliability in validity generalization: The consistency of Hermes, the utility of Midas. *Journal of Work and Organizational Psychology, 32*, 17-23. doi: 10.1016/j.rpto.2015.12.001

Salthouse, T. A. (1994). The aging of working memory. *Neuropsychology, 8*(4), 535–543. doi: 10.1037/0894-4105.8.4.535

Sattler, J. M. (2001). *Assessment of children: Cognitive applications* (4th ed.). San Diego, CA: Jerome M. Sattler Publisher.

Schaie, K. W. (1989). Perceptual speed in adulthood: cross-sectional and longitudinal studies. *Psychology and Aging, 4*(4), 443-453. doi: 10.1037/0882-7974.4.4.443

Schaie, K. W. (2005). *Developmental influences on adult intelligence: The Seattle Longitudinal Study*. New York: Oxford University Press.

Scharroo, J., & Leeuwenberg, E. (2000). Representation versus process in simplicity of serial pattern completion. *Cognitive Psychology*, *40*, 39–86. doi: 10.1006/cogp.1999.0722

Schmidt, F. L., & Hunter, J. (1998). The validity and utility of selection methods in personnel psychology: Practical and Theoretical Implications of 85 years of research findings. *Psychological Bulletin, 124*(2), 262-274. doi: 10.1037/0033-2909.124.2.262

Schmidt, F. L., & Hunter, J. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology, 86*(1), 162-173. doi: 10.1037/0022-3514.86.1.162

Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1981). Task difference and validity of aptitude tests in selection: A red herring. *Journal of Applied Psychology, 66*, 166–185. doi: 10.1037/0021-9010.66.2.166

Schmitt, N., & Klimoski, R. (1991). *Research methods in human resources management.* Cincinnati, OH: South-Western Publishing Co.

Schneider, J. W., & McGrew, K. S. (2012). The Cattell–Horn–Carroll model of intelligence. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 99-144). New York, NY: Guilford Press.

Shuttleworth, M. (2009). Test–Retest Reliability. *Explorable*. Retrieved from https://explorable.com/test-retest-reliability

Shye, S. (1988). Inductive and deductive reasoning: A structural reanalysis of ability tests. *Journal of Applied Psychology*, *73*(2), 308–311. doi: 10.1037/0021-9010.73.2.308

Siegler, R. S. (1992). The other Alfred Binet. *Developmental Psychology, 28*, 179-190. doi: 10.1037/0012-1649.28.2.179

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representations. In S. M. Downing & T. M. Haladyna, (Eds.), *Handbook of test development* (pp. 329-347). Mahwah, NJ: Lawrence Earlbaum Associates.

Snow, R. E., Kyllonen, C. P., & Marshalek, B. (1984). The topography of ability and learning correlations. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (pp. 47–103). Hillsdale, NJ: Erlbaum.

Spearman, C. (1904). General intelligence: Objectively determined and measured. *American Journal of Psychology, 15*, 201–293. doi: 10.2307/1412107

Spearman, C. (1914). The theory of two factors. *Psychological Review, 21*(2), 101-115. doi: 10.1037/h0070799

Spearman, C. (1927). *The nature of intelligence and the principles of cognition*. London: Macmillan and Co.

Spearman, C. (1938). *Measurement of intelligence*. Scientia, Milano.

Spellman, B. A., Holyoak, K. J., & Morrison, R. G. (2001). Analogical priming via semantic relations. *Memory & Cognition, 29*(3), 383-393. doi: 10.3758/BF03196389

Stankov, L. (2000). Complexity, metacognition, and fluid intelligence. *Intelligence, 28*(2), 121-143. doi: 10.1016/S0160-2896(99)00033-1

Sternberg, R. J. (1977). *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities.* Hillsdale, USA: Lawrence Erlbaum.

Sternberg, R. J. (1990). *Metaphors of mind: Conceptions of the nature of intelligence.* New York: Cambridge University Press.

Sternberg, R. J. (1999). The theory of successful intelligence. *Review of General Psychology, 3*, 292–316. doi:10.1037/1089-2680.3.4.292.

Sternberg, R. J. (2004). Culture and intelligence. *American Psychologist, 59*, 325-338. doi: 10.1037/0003-066X.59.5.325

Sternberg, R. J. (2005) The theory of successful intelligence. *Journal of Psychology, 39*(2), 189-202. doi: 10.1017/CBO9780511977244.026

Sternberg, R. J., Conway, B. E., Ketron, J. L., & Bernstein, M. (1981). People's conceptions of intelligence. *Journal of Personality and Social Psychology, 41*(1), 37. doi: 10.1037/0022-3514.41.1.37

Sternberg, R. J., & Detterman, D. K. (1986). *What is intelligence? Contemporary viewpoints on its nature and definition.* Norwood, NJ: Ablex.

Sternberg, R. J., Forsythe, G. B., Hedlund, J., Horvath, J. A., Wagner, R. K., Williams, W. M. (2000). *Practical intelligence in everyday life.* New York: Cambridge University Press.

Sternberg, R. J., & Gardner, M. (1983). Unities in inductive reasoning. *Journal of Experimental Psychology: General, 112*(1), 80–116. doi: 10.1037/0096-3445.112.1.80

Sternberg, R. J., Grigorenko, E. L., Ferrari, M., & Clinkenbeard, P. (1999). A triarchic analysis of an aptitude-treatment interaction. *European Journal of Psychological Assessment*, *15*(1), 1-11. doi: 10.1027//1015-5759.15.1.3

Thurlow, M., Lazarus, S. S., Albus, D., & Hodgson, J. (2010). *Computer-based testing: Practices and considerations* (Synthesis Report No. 78). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

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

Thurstone, L. L. (1947). *Multiple factor analysis.* Chicago, IL: University of Chicago Press.

Thurstone, L. L., & Thurstone, T. G. (1941). *Factorial studies of intelligence.* Chicago: University of Chicago Press.

Thurstone, T. G., & Thurstone, L. L. (1962). *Primary mental abilities tests.* Science Research Associates.

Tippins, N. T. (2009). Internet alternatives to traditional proctored testing: Where are we now? *Industrial and Organizational Psychology, 2*(1), 2–10. doi:10.1111/j.1754-9434.2008.01097.x.

Van de Vijver, F. J. R. (1991). *Inductive thinking across cultures: An empirical investigation.* Tilburg: Catholic Univ. of Brabant.

van den Noortgate, W., de Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics*, *28*, 369-386. doi: 10.3102/10769986028004369

van der Linden, W. J., & Glas, C. A. W. (2010). Preface. In W. van der Linden & C. Glas (Eds.), *Elements of adaptive testing* (pp. v–vii). New York, NY: Springer.

van der Maas, H. L., Kan, K. J., & Borsboom, D. (2014). Intelligence is what the intelligence test measures. Seriously. *Journal of Intelligence, 2*(1), 12-15. doi: 10.3390/jintelligence2010012

Verguts, T., & De Boeck, P. (2000). A note on the Martin-Löf test for unidimensionality. *Methods of Psychological Research Online, 5*, 77-82.

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

Visser, B. A., Ashton, M. C., Vernon, P. A. (2006). Beyond g: putting multiple intelligences to the test. *Intelligence 34*, 487–502. doi: 10.1016/j.intell.2006.02.004

Wainer, H. (2002). On the automatic generation of test items: Some whens, whys and hows. In S. H. Irvine, & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 287–316). New Jersey: Lawrence Erlbaum Associates.

Wechsler, D. (1939). *The Measurement of Adult Intelligence*. Baltimore: Williams and Wilkins.

Wechsler, D., Coalson, D. L., & Raiford, S. E. (2008). *WAIS-IV: Wechsler adult intelligence scale*. San Antonio, TX: Pearson.

Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences, 2*(1), 1-27. doi: 10.2458/azu_jmmss.v2i1.12351

Wharton, C. M., Holyoak, K. J., & Lange, T. E. (1996). Remote analogical reminding. *Memory & Cognition, 24*(5), 629-643. doi: 10.3758/BF03201088

White, S. (2000). Conceptual foundations of IQ testing. Psychology, *Public Policy, and Law*, *6*(1), 33-43. doi: 10.1037/1076-8971.6.1.33

Wilk, S. L., Desmarais, L. B., & Sackett, P. R. (1995). Gravitation to jobs commensurate with ability: Longitudinal and cross-sectional tests. *Journal of Applied Psychology, 80*, 79–85. doi: 10.1037/0021-9010.80.1.79

Wilk, S. L., & Sackett, P. R. (1996). Longitudinal analysis of ability-job complexity fit and job change. *Personnel Psychology, 49*, 937–967. doi: 10.1111/j.1744-6570.1996.tb02455.x

Wonderlic, E. F. (1983) *Wonderlic personnel test manual.* NorthField, IL: E. F. Wonderlic & Assoc.

Wonderlic, E. F. (1992). *Wonderlic personnel test user's manual.* Libertyville, IL: Wonderlic.

Wonderlic, E. F. (2007). *Wonderlic personnel test - revised* (WPT-R). Libertyville, IL: Wonderlic Personnel Test, Inc.

Zenisky, A. L., & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15*(4), 337-362. doi: 10.1207/S15324818AME1504_02

Zenisky, A. L., & Sireci, S. G. (2013, April 28-30). *Innovative items to measure higher-order thinking: Development and validity considerations*. Presented at the Annual Meeting of the NCME, San Francisco, CA.

APPENDIX

LETTER OF APPROVAL

LOUISIANA TECH
UNIVERSITY

MEMORANDUM

OFFICE OF UNIVERSITY RESEARCH

TO: Mr. Scott Hines and Dr. Tilman Sheets

FROM: Dr. Stan Napper, Vice President Research & Development

SUBJECT: HUMAN USE COMMITTEE REVIEW

DATE: August 3, 2016

In order to facilitate your project, an EXPEDITED REVIEW has been done for your proposed study entitled:

"The Development and Validation of an Automatic Item Generation Measure of Cognitive Ability"

HUC 1451

The proposed study's revised procedures were found to provide reasonable and adequate safeguards against possible risks involving human subjects. The information to be collected may be personal in nature or implication. Therefore, diligent care needs to be taken to protect the privacy of the participants and to assure that the data are kept confidential. Informed consent is a critical part of the research process. The subjects must be informed that their participation is voluntary. It is important that consent materials be presented in a language understandable to every participant. If you have participants in your study whose first language is not English, be sure that informed consent materials are adequately explained or translated. Since your reviewed project appears to do no damage to the participants, the Human Use Committee grants approval of the involvement of human subjects as outlined.

Projects should be renewed annually. ***This approval was finalized on August 3, 2016 and this project will need to receive a continuation review by the IRB if the project, including data analysis, continues beyond August 3, 2017.*** Any discrepancies in procedure or changes that have been made including approved changes should be noted in the review application. Projects involving NIH funds require annual education training to be documented. For more information regarding this, contact the Office of University Research.

You are requested to maintain written records of your procedures, data collected, and subjects involved. These records will need to be available upon request during the conduct of the study and retained by the university for three years after the conclusion of the study. If changes occur in recruiting of subjects, informed consent process or in your research protocol, or if unanticipated problems should arise it is the Researchers responsibility to notify the Office of Research or IRB in writing. The project should be discontinued until modifications can be reviewed and approved.

If you have any questions, please contact Dr. Mary Livingston at 257-2292 or 257-5066.