Beyond *g*: Assessing the Incremental Validity of the Cattell-Horn-Carroll (CHC) Broad Ability Factors on the Woodcock-Johnson III Tests of Cognitive Abilities

A Dissertation by

Ryan J. McGill

Chapman University

Orange, California

College of Educational Studies

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Education, with an emphasis in School Psychology

November 2013

Committee in charge:

Randy T. Busse, Ph.D., Chair

Kelly S. Kennedy, Ph. D.

John T. Brady, Ph. D.

Pedro Olvera, Psy.D.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

The dissertation of Ryan J. McGill is approved.

Randy T B

Randy T. Busse, Ph. D., Committee Chair

Kennedy

Kelly S. Kennedy, Ph. D.

John Brady

John T. Brady, Ph. D.

Pedro Olvera, Psy. D.

November 15, 2013

Beyond *g*: Assessing the Incremental Validity of the Cattell-Horn-Carroll (CHC) Broad Ability Factors on the Woodcock-Johnson III Tests of Cognitive Abilities

By

Ryan J. McGill

ACKNOWLEDGEMENTS

I would like to express my gratitude to the faculty and staff of the College of Educational Studies at Chapman University. It has been a privilege and an honor to have been able to spend the past four years in an environment that is dedicated to fostering the personal and intellectual development of its students. When I decided to enroll in the doctoral program at Chapman I had specific goals that I wanted to accomplish and I can truly say that this program has given me the opportunity and resources needed to work toward them all.

Specifically I would like to thank Randy Busse. For the past three years Dr. Busse has served as my advisor, dissertation chair, and personal guide into the world of academia and beyond. Thank you for the time and energy that you have dedicated to seeing me grow as an aspiring scholar. You provided me with my first opportunities to teach and conduct research and for that I am eternally grateful. I would also like to thank Kelly Kennedy for always having an open door for research talks and professional advice. Thank you additionally to the remaining members of my committee: John Brady and Pedro Olvera. Without your time and support I would not have been able to complete this journey.

I would also like to mention those that have provided technical and conceptual assistance with this project. First and foremost, Kevin McGrew and the Woodcock-Munõz Foundation for generously providing me with access to the dataset for this project. I would also like to thank Gary Canivez and Joe Glutting for technical assistance at various times during this project's conceptualization and completion.

On a personal level, a special commendation is reserved for James Milkovich. I will never forget what you and your family have done for me over the last decade. As a wise man once said, “I would follow you into the mists of Avalon.” I would also like to thank Minh Tran, who has waited patiently as I put my golf clubs aside for a long period of time in order to finish this project.

To all of the remaining family, friends, coaches, teachers, mentors, and colleagues both professional and academic whom I cannot name individually because of space limitations, you have all played a part in the journey that has led me to this point in my life. Thank you for all of your encouragement, motivation, and support throughout the years.

Last but not least, I would like to thank my wife Amber for her patience, understanding, support, love, and encouragement. No title, award, or initials at the end of my name will ever come close to bringing me the satisfaction that comes from being your husband.

ABSTRACT

Beyond *g*: Assessing the Incremental Validity of the Cattell-Horn-Carroll (CHC) Broad Ability Factors on the Woodcock-Johnson III Tests of Cognitive Abilities

By

Ryan J. McGill

Despite their widespread use, controversy remains about how to best interpret norm-referenced tests of cognitive ability. Due to the fact that contemporary cognitive measures appraise performance at multiple levels (e.g., subtest, factor, full-scale), a multitude of inferences about individual functioning are possible. Because school psychologists primarily utilize intelligence tests for predicting achievement outcomes, the cognitive variables that provide the most optimal weighting for prediction are of greatest importance. This study examined the predictive validity of the Cattell-Horn-Carroll (CHC) factor structure from the Woodcock-Johnson III Tests of Cognitive Abilities (WJ-COG; Woodcock, McGrew, & Mather, 2011c). Specifically, the incremental achievement variance accounted for by the CHC broad factors, after controlling for the effects of the General Intellectual Ability (GIA) composite, was assessed across reading, mathematics, writing, and oral language variables from the Woodcock-Johnson III Tests of Achievement (WJ-ACH; Woodcock, McGrew, & Mather, 2001b). Hierarchical regression was used to assess predictive relationships between the cognitive-achievement variables on the Woodcock-Johnson III assessment battery (WJ-III; Woodcock, McGrew, & Mather, 2001a). This study utilized archived standard score data from individuals (*N* = 4,722) who participated in the original WJ-III standardization project. Results showed that the GIA accounted for the largest portions of achievement for all but one of the regression models that were assessed. Across the models, the GIA variance coefficients represented moderate to large effects whereas the CHC factors accounted for non-significant incremental effects in most of the models. Nevertheless, the WJ-COG factor scores did account for meaningful portions of achievement variance in several situations: (a) in predicting oral expression scores; (b) in the presence of significant inter-factor variability; and (c) when the effects of Spearman's law of diminishing returns (SLODR) was accounted for in reading, mathematics, and written language regression models. Additionally, the chi-square goodness of fit test was utilized to assess model invariance across several moderating variables. Results suggest that incremental validity is not a unitary construct and is not invariant across samples on the WJ-COG. Additionally, simultaneous interpretation of both the GIA and CHC factor scores on the WJ-COG may be useful within specific clinical contexts.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter I: Introduction

When children present with academic difficulties at school they are often referred to a school psychologist for the purposes of diagnosing and remediating the cause of the presenting problem. This process involves making inferences about the impact of psychological variables on a child's potential to benefit from instruction. Such an appraisal frequently involves the administration and interpretation of a comprehensive measure of intelligence. The use of intelligence tests for such purposes is based upon the long-held theory that specific cognitive processing deficits are the primary causes of severe academic difficulties such as specific learning disability (SLD; Fletcher, Lyon, Fuchs, & Barnes, 2007), and that cognitive variables reliably predict or account for a substantial amount of achievement variance as reflected across various reference markers (e.g., norm-referenced achievement tests). Despite many references to a "paradigm shift" in the field toward a problem-solving model of service delivery (e.g., Reschly, 2008), school psychologists continue to report administering and interpreting intelligence tests more frequently than any other measure in SLD assessments (Fagan & Wise, 2007). As such, the use and interpretation of intelligence tests continues to be a foremost concern within the field of school psychology.

Evolution of Intelligence Test Theory and Interpretation

Despite their widespread use, a long-standing controversy remains about how to best interpret cognitive measures. Interpretation of intelligence tests involves making inferences about an individual based upon the scores that are obtained from a standardized administration of a particular test. Due to the fact that contemporary cognitive measures appraise performance at multiple levels (e.g., subtest, composite, full-scale); practitioners potentially make a multitude of inferences regarding individual performance. Because intelligence tests primarily are utilized for predicting important outcomes such as achievement, the cognitive variables which provide the optimal weighting for prediction are of greatest importance.

Initially, intelligence test interpretation was limited to evaluating individual performance at the full-score level. This practice was due to the fact that most early intelligence tests were modeled according to Spearman's "*g*-theory" and only provided users with a full-scale composite score of overall general intelligence. Buoyed by the development of more robust methods for factor analytic modeling, subsequent versions of tests provided users with a variety of subordinate composite score options that measured several common factors such as visual processing, perceptual reasoning, verbal skills, and processing speed, as a complement to the full-scale score. During the second half of the 20th century, a variety of "clinical" methods were proposed (e.g., Kaufman, 1994) that emphasized the sequential interpretation of individual cognitive strengths and weaknesses as reflected in subtest and index score performance. A dominant feature of such methods was the de-emphasis of the importance of *g* and its corresponding full-scale score.

Over the last two decades numerous psychometric investigations (e.g., Glutting, Watkins, & Youngstrom, 2003; Watkins, 2000) have been conducted to assess the reliability and validity of the proposed clinical interpretation methods. As a result of these investigations, these methods (e.g., "intelligent testing") were found to lack the requisite reliability or validity for clinical decision-making. These studies resulted in a broad consensus within the field of school psychology that subtest level analysis should not be used for individual decision-making (Watkins, Glutting, & Youngstrom, 2005).

However, the use of profile analysis with higher-level factor scores continues to be a mainstay of contemporary clinical practice.

Recently, individual tests have been constructed to reflect various theories of intellectual functioning, and interpretations may be based upon the theory on which the test is based (Canivez, 2013b). Although there has always been a symbiotic relationship between intelligence theory and the science of psychological test development and interpretation known as psychometrics, this relationship is the highlight of recently proposed methods for interpreting intelligence test data that emphasize the role of cognitive and neuropsychological theory in appraising individual differences (Kamphaus, Winsor, Rowe, & Kim, 2012). The most prominent of these theories, known as the Cattell-Horn-Carroll theory (CHC; McGrew, 2005), has come to dominate the contemporary intelligence testing landscape and is considered by some, in the academic literature (e.g., Burns, 1994; McGrew, 2009), to be the most comprehensive and empirically validated theory of cognitive abilities.

Intelligence testing research has changed dramatically over the last two decades as a result of the widespread incorporation of CHC theory in the school psychology canon. CHC theory has been used to provide a theoretical and empirical foundation for understanding cognitive abilities and how they relate to academic learning, which has had a profound impact on school psychology clinical practice (Alfonso, Flanagan, & Radwin, 2005). CHC theory is based upon a synthesis of several theories of intellectual functioning, including Spearman’s two-factor theory (Spearman, 1904), Thurstone’s primary mental abilities (Thurstone, 1938), Cattell and Horn’s Gf-Gc theory (Horn & Cattell, 1966), and Carroll’s three-stratum theory (Carroll, 1993). In the CHC model, various cognitive abilities are distributed along a three-stratum structure with stratum I representing “narrow” abilities that feed into one of seven stratum II “broad” common factors. The common factors all load onto a higher-order *g*-factor that is thought to represent overall intellectual ability at stratum III. In the last decade a number of authors of book chapters, monographs, and scientific papers have argued for the reorganization of most conventional intelligence tests using the CHC model as a blueprint. Although several tests reference CHC theory within their manuals (e.g., Kaufman Assessment Battery for Children-Second Edition; KABC-II), the Woodcock-Johnson III Tests of Cognitive Abilities (WJ-COG; Woodcock, McGrew, & Mather, 2001c) is the only current test to utilize CHC specifically as its foundation, and to as measure all of the proposed broad cognitive abilities within the CHC model. Thus, it serves as a valuable research tool for conducting investigations related to better understanding human cognitive abilities.

To fill the void that was created by the demise of the traditional discrepancy model of SLD identification, a number of alternative models (e.g., Flanagan, 2003) have been tendered which call for clinicians to analyze profiles of individual cognitive strengths and weaknesses derived via CHC part-scores. Although several such models have been proposed within the last five years, they all fall under the general umbrella of an approach to SLD identification that is known as “profile of strengths and weaknesses” or PSW. The substance of the PSW model is based upon the accumulation of a long program of research (see McGrew & Wendling, 2010), which has demonstrated empirical links between specific CHC cognitive variables and individual areas of achievement. With continued emphasis on the assessment of cognitive processes, it is critical for practitioners to have a strong theoretical and empirical foundation when making high-stakes educational decisions.

Incremental Validity and Hierarchical Multiple Regression

To establish the psychometric integrity of subordinate part-scores on intelligence tests, it is critical to examine relationships between such scores and external criteria such as achievement indicators. Conventional intelligence tests such as the WJ-COG are hierarchically structured with the weighted combination of factors utilized to create a full-scale composite therefore, external validity investigations which assess the incremental and predictive validity of lower-level scores beyond that of higher-level scores is critical (Sechrest, 1963). The purpose of incremental validity studies is to demonstrate which variables account for relevant variance in a domain of interest (e.g., achievement) beyond that already accounted for by existing variables or test data. As applied to intelligence tests, the concern is with the additional achievement variance accounted for by subordinate scores (e.g., factor, subtest) beyond that provided by the full-scale composite or IQ score.

As previously noted, several empirical studies have demonstrated relationships between specific WJ-COG broad cognitive factors and reading (Floyd, Keith, Taub, & McGrew, 2007), mathematics (Taub, Keith, Floyd, & McGrew, 2008), and writing (Floyd, McGrew, & Evans, 2008) utilizing a multivariate statistical method for assessing latent variables known as structural equation modeling (SEM). However, the results of SEM studies cannot be used as evidence of the incremental validity of factors for several reasons, namely: SEM assesses variables at the latent level which do not have direct implications to practitioners who work with variables at the observed level, and SEM is predominately utilized to assess hypothesized model fit, thus it is a poor method for variance partitioning.

A more robust method for partitioning variance, popularized by Cohen and Cohen (1983), is sequential or hierarchical multiple regression (HMR). In HMR, the dependent variable variance accounted for by one or more independent variables is assessed while holding constant the proportion of variance already accounted for by an additional prescribed independent variable. The amount of additional variance that is provided by the other independent variable(s) can then be evaluated statistically through effect size estimates and statistical tests of significance. It should be noted that order or entry is an issue in HMR designs, thus the decision as to which variable to control for should be based upon a formulated theory (Pedhazur, 1997).

Using an example with conventional intelligence test variables, a typical HMR design might utilize various achievement variables in the form of standard scores from a norm-referenced test of achievement as dependent variables in separate analysis, with the full-scale score entered into the first block of the regression equation with factor scores entered into the second block. By entering the factor scores into the regression equation in the second block, the predictive variance that is accounted for by the full-scale score is controlled for or partitioned out from the proportion attributable to the factor scores. Authors such as Hale, Fiorello, Kavanaugh, Hoeppner, and Gaither (2001) have argued that it is possible in the above example to just as easily enter the factor scores in the first block and the full-scale score into the second. However, their argument is inconsistent with existing intelligence theory as most models posit a higher order *g*-factor that is principal to lower-order factors. Utilizing a design in which part scores are entered into the regression equation prior to the full-scale score is akin to putting the cart before the horse (Schneider, 2008).

HMR has been extensively utilized to assess the incremental validity of several intelligence tests, including the Wechsler Intelligence Scale for Children-Fourth Edition (Glutting, Watkins, Konold, & McDermott, 2006), KABC-II (McGill & Busse, 2012), Reynolds Intellectual Assessment Scales (Nelson & Canivez, 2012), and the Cognitive Assessment System (Canivez, 2011). A synthesis of these studies indicates that across samples, factor scores account for negligible amounts of incremental achievement prediction beyond that already accounted for by the full-scale score (Canivez, 2013b). However, a comprehensive incremental validity investigation on a contemporary intelligence test founded primarily upon CHC theory such as the WJ-COG has yet to be conducted.

Purpose of the Current Study

This study is designed to assess four research areas. First, the overall incremental validity of the WJ-COG will be assessed using several areas of achievement as measured by the Woodcock-Johnson III Tests of Achievement (WJ-ACH; Woodcock, McGrew, & Mather, 2001b) as dependent variables in an HMR design. This research is important because it will provide a foundational basis for establishing the construct validity of the CHC broad factors that are measured on the WJ-COG. The results of this investigation will have implications that are potentially relevant for evaluating the WJ-COG interpretive framework, the diagnostic validity of proposed PSW models of SLD identification, and the external validity of the broader CHC framework.

The second purpose is to assess whether the incremental validity of the CHC factors are invariant across levels of schooling. The third purpose of this study is to determine whether the results obtained from the initial incremental validity investigation are invariant across several levels of inter-factor variability. In other words, do significant differences between the crystallized ability (Gc) and fluid reasoning (Gf) factors have an impact on the predictive validity of the full-scale score composite? This research question is important due to the fact that several authors of popular interpretive textbooks (e.g., Flanagan & Kaufman, 2004; Hale & Fiorello, 2004) have argued that full-scale scores on intelligence tests should not be interpreted whenever significant inter-factor variability is observed. Such an approach is also advocated in many test manuals and is a widely utilized interpretive technique in clinical practice (Pfeiffer, Reddy, Kletzel, Schmelzer, & Boyer, 2000).

The remaining purpose of this study is to assess whether the phenomenon known as Spearman's Law of Diminishing Returns (SLODR) has an impact on the incremental validity of the WJ-COG. Briefly, SLODR is based upon research conducted by Spearman (1927) in which he observed that the correlation or positive manifold between cognitive tests was greater for individuals with lower levels of estimated ability than for those who had higher levels of overall ability. The implication of this phenomenon is that as overall ability rises, the *g*-factor becomes less ubiquitous. The results of this investigation will provide researchers and practitioners with additional information as to the potential effects of SLODR on intelligence test interpretation.

This study is utilizing the WJ-COG because it is one of the most popular intelligence tests utilized in contemporary practice by school psychologists, has satisfactory psychometric properties for use in research and clinical settings (Sandoval, 2003), and is the only intelligence test that has been founded solely on the basis of CHC theory. As previously indicated, CHC theory is one of the most empirically validated theories of intelligence found within the academic literature (McGrew, 2009).

Summary

As school psychologists predominately utilize intelligence tests in a predictive manner (e.g., to account for observed achievement variance), an improved understanding of the optimal weighting scheme for various cognitive test variables in prediction equations is critical for evidence-based practice. Incremental validity investigations have potential implications for the clinical interpretation of intelligence test measures, the validity of PSW models, and the external validity of the broader CHC model. This study is an examination of the incremental validity of the CHC broad cognitive factors on the WJ-COG while accounting for the potential impacts of maturation, inter-factor variability, and the effects of SLODR on the observed results.

Chapter II: Literature Review

The purpose of this section is to provide a comprehensive review of the literature relevant to the current study. The topics that compose this review include a history of intelligence research leading up to the synthesis of Cattell-Horn-Carroll theory, a review of methods for interpreting intelligence tests, and an overview of incremental validity and the various methods that have been utilized to assess it with current intelligence tests. The chapter concludes with an overview and rationale for the current study.

Literature Search

The literature included in this chapter was located by conducting a search of books, monographs, technical papers, and journal articles published between 1855 and 2013, utilizing the Educational Resources Information Center (ERIC), PsycARTICLES, and PsycINFO databases using a combination of the following descriptors: cognitive ability, intelligence test, incremental validity, CHC, and achievement outcomes. Manual searches also were performed in the following journals: *School Psychology Review, School Psychology Quarterly, Journal of School Psychology, Journal of Applied School Psychology, Psychology in the Schools, Contemporary School Psychology, Journal of Psychoeducational Assessment, Intelligence, Psychological Assessment, Journal of Special Education,* and *Applied Neuropsychology*. The reference sections of individual articles also were searched to identify additional primary and secondary sources of relevance to the present study. Additionally, a digital search, using a combination of the descriptors: *incremental validity*, *predictive validity*, *CHC*, and *test*, was completed using the ProQuest dissertations and theses database to locate relevant non-published empirical studies. Finally, several individual researchers were contacted directly to inquire about the existence of any non-published data that might be of relevance to the current study.

History of Intelligence Theory and the Development of Intelligence Testing

Because this study is concerned with assessing the validity of CHC factors, it is necessary to trace the evolution of intelligence theory, which has culminated in the development of the Cattell-Horn-Carroll (CHC; McGrew, 2005) model of intelligence. Such an exercise is important for several reasons, the most important being the fact that most contemporary intelligence tests now reference CHC within their theoretical and interpretive frameworks. A review of previous theory and interpretive methods is necessary to provide historical context for contemporary debates within the field of school psychology that are relevant to understanding how such measures should be interpreted. As Boring wrote, “without such knowledge he sees the present in distorted perspective, he mistakes old facts for new, and he remains unable to evaluate the significance of new movements and methods” (1929, p. vii).

Although a host of sociological factors have influenced the development of intelligence theories and the tests that have been proposed to measure cognitive abilities (Wasserman, 2012), the primary focus of this chapter is limited to understanding the development of intelligence theory and testing practices from a psychometric perspective. Therefore, intellectual models that are primarily theoretical, cognitive, or socio-cultural in nature (e.g., multiple intelligence theory, information processing, triarchic model of intelligence) are not included in this review. This chapter continues with a synthesis of several psychometric models of intelligence and how they have influenced the development of current intelligence tests.

Early Developments

Despite its contemporary importance, few psychology textbooks published in the latter half of the 19th century referenced the construct of intelligence. This is not to say that society at that time did not place value in the appraisal of individual differences, as efforts to identify and reward individuals with “superior” ability have been recorded for several centuries (Suen & French, 2003). As an example, China has employed a large scale civil service examination system continuously since the 1300s (Bowman, 1989), many years prior to the introduction of comparable systems in western civilization.

The construct of intelligence is derived from the field of differential psychology, which is largely concerned with the study of biological and quantitative differences in behavior amongst individuals. The intellectual roots of the field can be traced to philosopher Herbert Spencer. Although Spencer is most famous for coining the term “survival of the fittest” after being influenced by the evolutionary writings of biologist Charles Darwin, he had previously written a textbook of psychology in which he defined intelligence as the “continuous adjustment of inner to outer relations” (1855, p.486). Jensen (1998) noted that a thorough history of the development of intelligence testing is incomplete without mentioning the influence of Darwin and Spencer, as their writings served as a major influence on the development of early empirical psychology and its search for natural laws that governed behavior across the species.

The empirical study of mental ability did not begin until the early “brass instrument” experiments that were conducted in England by Francis Galton from 1884 to 1890. Galton designed a battery of cognitive skills largely composed of elementary sensory discrimination tasks with the belief that such measures were largely unaffected by social standing and prior education. He also believed that perception served as a conduit to the latent construct of intellect: “The only information that reaches us concerning outward events appears to pass through the avenue of our senses; and the more perceptive the senses are to difference, the larger the field upon which our judgment and intelligence can act” (Galton, 1883, p. 27). However, the narrow focus on the measurement of primary senses inhibited refinement and understanding of many higher-order cognitive processes (e.g., working memory) which are now commonly referenced within the contemporary intelligence literature. Despite this limitation, Galton’s contribution to differential psychology cannot be overstated as his experiments constituted the first standardized testing program and served as a major influence in the development of more refined assessment technologies.

Binet-Simon Scale

An important moment in the history of IQ testing occurred in 1904 when the Minster of Public Instruction in Paris appointed Alfred Binet to a committee whose charge was to develop a methodology for accurately identifying children with significant intellectual challenges in the general population. Prior to his inclusion on the commission, Binet had been working on the development of an objective and standardized assessment system for diagnosing childhood educational difficulties (Anastasi & Urbina, 1997). Along with his colleague Theodore Simon, he developed the Binet-Simon scale, an assessment system that was composed of several interviews and a 30-item cognitive battery. In contrast to previous cognitive measures, the Binet-Simon scale included age-based items designed to account for the changes that occur in cognitive growth during the course of development. Although Binet was influenced by the work of Galton, he recognized the limitations of focusing solely on sensory discrimination tasks and incorporated measures of higher-order processes such as memory span and verbal comprehension within his battery.

The 1905 scale functioned more as a clinical interviewing tool rather than a traditional standardized test. It did not provide a summary measure of intellectual ability. A revision of the scale in 1908 provided users with a summary statistic referred to as a mental age, a forerunner to the modern day intelligence quotient. Commercial success eluded Binet in his native France though his contributions were later reified with the introduction of the Binet-Simon scales in the United States by Henry Goddard and Stanford psychologist Lewis Terman. Although Goddard was one of the first American psychologists to utilize previous versions of the scales in his Vineland Training School, Binet sold the commercial rights to publish his test in the United States to Terman who subsequently renamed the American version the Stanford-Binet scale (Wolf, 1973).

Along with the Wechsler scales, the Stanford-Binet dominated the cognitive testing landscape in the U.S. for the better part of the 20th century (Becker, 2003). Although the original Binet-Simon scale is considered rudimentary according to modern test development standards, the development of the first usable measure of cognitive ability is considered to be one of the great accomplishments in the field of psychology (Thorndike & Lohman, 1990).

Although the 1904 Paris committee was largely focused on the identification of children with intellectual disabilities, the new scale was largely utilized as a tool to predict future academic performance. Thus, since their inception, one of the more practical purposes of intelligence scales has been their utility in predicting socially significant outcomes (Gottfredson, 1997). Attempts to explain why his test worked so well were largely limited to intuition and conjecture though in remarkably keen insight for the times, Binet and Henri wrote in an 1895 *L'Annee Psychologique* manuscript: "the higher and more complex a process is, the more it varies in individuals…if one wishes to study the differences existing between two individuals, it is necessary to begin with the most intellectual and complex processes, and it is only secondarily important to consider the simple and elementary processes" (as cited in Wasserman, 2012). The ability to demonstrate this phenomenon empirically would soon be possible with the development of the field of psychometrics and the work of Charles Spearman.

Spearman's Two Factor Theory

Buoyed by the statistical developments of his cousin Karl Pearson, British statistician Charles Spearman proposed the first unified theory of cognitive abilities with the publication of the paper "General Intelligence, Objectively Determined and Measured" in 1904. Although Spearman did not label his theory, it is often referred to in the contemporary literature as the "*g* theory." Although Spearman conducted his experiments around the same time as the development of Binet's scale, there appears to have been limited professional interaction between the two.

Spearman (1904) administered a battery of assessments to a sample of 60 school-aged children in a small village in England. The tests measured subjects such as classics, French, English, math, vocal pitch, and sound discrimination. He determined that all of the measures tended to correlate, a phenomenon later referred to as positive manifold (Thurstone, 1947). Spearman then arranged all of the coefficients between the tests into a matrix which he then analyzed using a primitive form of factor analysis known as the method of tetrad differences. In his analysis he found that 62.9% of the total variance between all of the tests was accounted for by a single factor which he identified as the general factor or *g*. The remaining 37.1% of variance was attributed to specific factors unique to the individual tests themselves, a factor he identified as *s*.

Spearman appeared to remain ambivalent about the exact nature of *s*, whose influence he stated was largely negated by the combination of individual test scores into larger composites. Despite many criticisms, he appeared to resist the notion of including additional common factors in his model because he stated that it would open the door for the inclusion of an infinite number of hypothesized subordinate factors; though by the end of his career he stipulated with many of his theory's critics that cognitive ability may be better represented by a second order *g* factor with an undetermined number of first-order common factors which represented more discrete cognitive skills (Spearman & Jones, 1950).

Spearman's hypothesis postulates that all tests of cognitive ability are composed of some form of *g* variance, an observation he referred to as the "indifference of the indicator" (Spearman, 1927, p. 197). A clear distinction must be made between *g* and the theoretical entity of intelligence. Whereas there appears to be a relationship between some features of intelligence and psychometric *g*, it is unlikely that all of what is known regarding intelligent behavior can be reduced to a summary statistic. Regarding the subject Terman wrote: "We must guard against defining intelligence solely in terms of ability to pass the tests of a given intelligence scale. It should go without saying that existing scales are incapable of adequately measuring the ability to deal with all possible kinds of material on all intelligence levels" (1921, p. 131). Those words remain as important to scholars of intellectual ability today as they were when Terman first penned them.

Although the ubiquity of the *g*-factor is well supported within the empirical literature, less is known about some of the more technical aspects of *g*. Subsequent factor analytic research has demonstrated that *g* is not invariant across various clinical conditions. Recent research has shed some light on several factors which impact the degree to which various cognitive measures load on the general factor, though it is worth reiterating that all cognitive tasks demonstrate some degree of *g* loading, which is identified in studies of cognitive abilities as common variance.

In general, *g* loadings are more robust in tasks that are more cognitively complex (Brody, 1992). Variance partitions that have been conducted on subtests from multidimensional ability batteries have consistently revealed that tasks that require more complex forms of mental processing such as inductive and deductive reasoning, mental rotation, working memory, and verbal comprehension are composed of higher *g* loadings than tasks that involve more automatic forms of processing such as psychomotor speed, visual processing, and memory span (Sattler, 2008). To illustrate, in a recent study of the Differential Ability Scales-Second Edition (DAS-II; Elliott, 2007) conducted by Maynard, Floyd, Acklie, and Houston (2011) using a sample of 3,106 school-aged children and adolescents, average general factor variance contained within the crystallized and fluid ability composite scores ranged from 51% to 60%, whereas the processing speed index loading was 25% (in contrast to 64% specific variance). These results are consistent with those that have been obtained on other intelligence tests that measure the same constructs (e.g., Floyd, Bergeron, McCormack, Anderson, & Hargrove-Owens, 2005).

Another factor that has been shown to impact *g* loadings is a phenomenon known as Spearman's law of diminishing returns (SLODR). Spearman (1927) compared correlation matrices between 78 "normal" children and 22 "defective" children and found that the mean correlation coefficient between tasks was .47 for the normal sample and .78 for the sample of children with below average intellectual abilities. From these data he concluded that "the more 'energy' a person has available already, the less advantage accrues to his ability from further increments of it" (p. 219).

SLODR is less understood then other factors influencing *g* due to the fact that it has only recently begun to be studied within the scientific literature. The first large scale empirical study (*N* = 4,080) of the effects of SLODR was conducted by Detterman and Daniel (1989). They divided a sample of individuals who had been administered the Wechsler intelligence scales into five ability groups and found that inter-correlations between subtests decreased across the low to high ability groups. In a more sophisticated large-scale study (*N* = 10,535), Deary and colleagues (1996) demonstrated differential *g* loadings across ability groups along with as evidence for an interaction effect across ages.

Several recent investigations using factor mixture modeling have provided latent variable evidence for differential loadings across ability groups on contemporary intelligence tests (Reynolds, Hajovsky, Niileksela, & Keith, 2011; Reynolds, Keith, & Beretvas, 2010). On the basis of such investigations, Reynolds (2013) encouraged practitioners to interpret broad ability profiles in place of general factor scores for individuals with higher estimated levels of general intelligence. However, little is known about the effect that SLODR may have on the predictive validity of IQ tests with observed-level variables. A recent investigation using a combination of simulated and observed cognitive test data ($N = 436$) by Murray, Dixon, and Johnson (2013), questioned the empirical support for SLODR, concluding that it was an artifact of method variance.

Despite the substantial validity evidence for the *g*-factor, there are detractors as to the practical applications of the construct. For example, Valencia and Suzuki (2001) encouraged practitioners to remain agnostic about the very existence of *g* until there is consensus regarding its definition. Others (e.g., Sternberg, 1984) posit that various cultures define intelligence somewhat differently; therefore the construct itself (as represented by *g*) may be nothing more than an artifact. However, Gottfredson (2005a) stated that such arguments are overly restrictive and serve to promote a form of intellectual nihilism. It may be unlikely that those who disagree will ever be persuaded regarding the relative merits of *g* (or the use of intelligence tests for that matter). As Schneider (2011) wrote, “If you think the matter will be settled by accumulating more data, you have not been paying attention for the last hundred years.” Given the criticism surrounding the construct it is tempting to disregard the *g*-factor entirely. However, proponents of such an approach should take into consideration the practical validity evidence for *g* and its predictive relationship with important life outcomes (e.g., Herrnstein & Murray, 1994; Schmidt & Hunter, 2004).

Before proceeding it is important to differentiate between common factors and *s*. Spearman’s model left little room for additional factors between the extremes of complete generality and complete specificity. A common factor (also known as a group factor) is identified when select tests within a battery share variance with each other that is more general than *s* but not prevalent enough to be considered a rival to the *g* factor. The debate regarding the validity of group factors has been one of the more contentious issues in all of differential psychology, serving as an impetus for the systematic development of several comprehensive models of cognitive abilities.

The Rise of Multiple Factor Theories

One of Spearman's critics was the Columbia University psychologist Edward L. Thorndike. Thorndike developed and facilitated the administration of a test to 63 primary and secondary school students that purported to measure several psychoeducational abilities such as sensory discrimination, quantitative reasoning, and vocabulary development. After reviewing correlational data, he concluded in a 1909 paper that "there is nothing whatsoever common to all mental functions, or even half of them" (Thorndike, Lay, & Dean, 1909, p. 368). Thorndike apparently rejected the notion of a general intelligence factor in favor of a model that emphasized multiple faculties of the mind.

Although Spearman spent the latter part of his career defending his findings from Thorndike's response, advancements in research methodology and statistical techniques allowed for the discovery of group abilities in cognitive assessment data. Spearman acknowledged these findings when specific cognitive tasks were found to load on group factors subordinate to *g*. These discoveries helped pave the road for the development of more empirically derived theories of mental ability.

L. L. Thurstone (1938) later developed a model of mental ability that was derived from a statistical technique that allowed for factors to be extracted from an extant data set, a method known as factor analysis. Using data from a battery of 56 mental tests that were administered to 240 college students he extracted seven factors which he described as visual-spatial, perception of visual detail, numerical, verbal logic, verbal words, memory, and induction. He called these factors “primary mental abilities” (PMA), a term which soon became associated with his model of intelligence. Thurstone later reconfigured his model to account for eight primary abilities. Although Thurstone later accepted the existence of a general factor, he stated that the use of a single score to estimate overall mental ability was inadequate for clinical decision making, and he encouraged the synthesis of an individual’s profile of scores across several measures of cognitive functioning to determine individual cognitive strengths and weaknesses.

The influence of Thurstone’s work on the modern-day understanding of the structure of human cognitive abilities cannot be overstated, as many of the group factors that he identified later served as the foundation for subsequent models of intelligence (e.g., CHC), alhough his apparent reluctance to embrace a general factor may be an artifact of the methods he used to identify his group factors. Thurstone utilized rotation techniques in his factor analyses that left the various broad abilities orthogonal (not correlated) to each other. By using such methods it was almost impossible for a general factor to be derived because little common variance in the factors could be extracted. Researchers using Thurstone’s PMA data sets later extracted a *g*-factor using oblique rotations; factor analytic techniques that assumed his broad abilities were indeed correlated (Jensen, 1998).

Influenced by Thurstone’s work, J. P. Guilford established the Aptitudes Research Project at the University of Southern California and conducted a number of studies from 1949 to 1969 with the purpose of identifying and organizing additional primary factors.

Using centroid methods, Guilford (1967) identified dozens of additional factors which he organized into an alternative taxonomic structure which he called the Structure of Intellect (SOI). However, independent scholars were unable to replicate many portions of the SOI model and were highly critical of the methods used to identify the various SOI factors (Carroll, 1993). Whereas support was found for some of the features identified, the broader model was rejected ultimately by a majority of the scientific community.

Nevertheless, Guilford deserves praise for being an early pioneer of what is now known as facet modeling, which differs significantly from traditional psychometric assessment approaches. In a facet model, factors are organized according to both their structural and functional properties. Whereas a more extensive discussion regarding facet theory is beyond the scope of this review, it is worth noting that variations of this approach have been utilized successfully in more recent investigations to model intellectual functioning (see Guttman & Levy, 1991; Snow, Kyllonen, & Marshalek, 1984).

Fluid-Crystallized Theory

A dichotomous model of ability followed known as the fluid and crystallized model of intelligence, or Gf-Gc theory. In a commentary discussing issues unresolved in the measurement of adult intelligence, Cattell proposed that cognitive ability was best represented by the presence of two general factors (1943), which he identified as fluid intelligence (g_f) and crystallized intelligence ($\mathrm{g_c}$). The convention was later adapted to reference group factors with an uppercase 'G' and the general factor by a lower case italicized '*g*'. These guidelines continue to be utilized today and have become standard practice for reporting empirical results involving CHC theory.

Influenced by Hebb's (1942) distinction between intelligence derived from biology and that derived from prior education, Cattell described fluid ability as the general facility in reasoning wherein prior knowledge cannot be utilized to solve problems, in contrast to crystallized ability which refers to the storage, retrieval, and use of prior knowledge. The original Gf-Gc theory provided an organizing framework for group abilities and provided many theoretical explanations for observations that had long puzzled researchers.

Gf-Gc was largely neglected for two decades before Cattell (1963) conducted the first experimental analysis of the theory by administering a series of nine cognitive tasks to a sample of school-aged children (*N* = 277) and subjecting the results to a factor analysis. His analysis indicated that each of the tasks primarily loaded on one of the two "intelligence" factors. Interestingly, he chose not to include a general factor in his model despite the fact that he stipulated that Gf and Gc were highly correlated and that a third-order factor solution was tenable. On the subject of Spearman's *g*, Cattell posited that *g* operated largely through Gf and he later developed a vehicle for describing interactions between Gf and Gc, which later became known as investment theory.

According to Cattell (1987), fluid ability serves as a limiting factor in how much information individuals can acquire from the environment. Therefore, learning is a function of the interaction between inherited levels of fluid ability and interpersonal meta-cognitive factors (e.g., motivation, drive, personality) which regulate how much of that fluid ability is invested by the individual within their environment. The product of that investment is later expressed in the form of developed crystallized ability. Cattell stated that this interaction helped explain why Gf and Gc were so highly correlated.

The first replication of Gf-Gc theory was conducted by John Horn in his doctoral dissertation at the University of Illinois. Horn (1965) administered 31 cognitive and personality trait tasks to a sample of 297 adults. He extracted several second-order factors from the data which he identified as fluid intelligence (Gf), crystallized intelligence (Gc), general visualization (Gv), general speediness (Gs), facility (F, a forerunner to long-term retrieval), carefulness (C, general cognitive accuracy), premsia (PRM, literary and artistic references), and positive self-image (PSI). Horn extracted two general factors which he did not further identify. The first general factor was composed of Gf, Gv, Gs, and F and the second was composed primarily of Gc, Gf, and PSI.

From 1966 through the late 1990s, Horn and Cattell collaborated in a systematic program of research aimed at validating and adding additional second-order factors to the Gf-Gc model. By the early 1990s, the Gf-Gc model had expanded to include nine broad second-order abilities: fluid intelligence (Gf), crystallized intelligence (Gc), short-term acquisition (Gsm), visual intelligence (Gv), auditory intelligence (Ga), long-term storage and retrieval (Glr), cognitive processing speed (Gs), correct decision speed (CDS), and quantitative knowledge (Gq). Additionally, Woodcock (1990) proposed the inclusion of a reading and writing ability factor (Grw).

One of the more consequential discoveries of the Gf-Gc research program has been the demonstration of differential declines in various broad abilities over the course of the human lifespan. In general, it has been demonstrated that Gc has been shown to increase through adulthood with small declines emerging around age 70 and beyond (Ackerman, 1996). Conversely, Gf skills have been shown to peak around early adulthood (i.e., ages 25 to 30) and then decline throughout the rest of the lifespan

(Verhaeghen & Salthouse, 1997). Based upon regression growth models, Noll and Horn (1998) estimated that the loss in Gf ability in adulthood was equivalent to 0.5 to 1.0 IQ units per decade. McGrew and Woodcock (2001) later argued that in spite of strong Gf-Gc correlations, such developmental validity evidence demonstrated that Gf and Gc are in fact orthogonal, unrelated abilities.

Development of Hierarchical Solutions

Philip Vernon (1950) is credited with articulating the first hierarchical model of cognitive abilities in which he posited that a higher-order *g* factor presides over two lower-order factors which he identified as verbal ability and spatial ability. The lower-order factors are composed of dozens of narrow abilities such as psychomotor coordination, attention, fluency, reasoning, and reaction time. Vernon stated that the model was most likely under-identified, and hypothesized that additional group factors beyond verbal reasoning and spatial thinking constituted a more complete model of cognitive ability. Although his model was never reconstituted, it provided empirical support for the verbal-nonverbal dichotomy of cognitive abilities that was popular among clinicians as a result of the publication of the Wechsler intelligence scales. Nevertheless, Vernon's model was seen as an important reconciliation between Spearman's two-factor model and Thurstone's primary abilities.

A more direct hierarchical test of the nature of cognitive abilities was conducted by Gustafsson (1984), who administered a battery of 16 tests to 1,000 sixth grade students and utilized factor analysis to test the fit of several competing models. He suggested that the model that best fit the data was a third-order *g* factor which reigned over three group ability factors. He referred to this model as the hierarchical LISREL model or HILI. Interestingly, Gustafsson found that the fluid reasoning factor was nearly identical to the third-order general ability factor, a finding which has served to perpetuate the theory within the scientific community that fluid reasoning is largely a proxy for the *g*-factor.

A major breakthrough in applied psychometrics occurred with the publication of *Human Cognitive Abilities: A Survey of Factor-Analytic Studies* (Carroll, 1993). Carroll assembled a collection of over 400 datasets of factor-analytic studies of cognitive abilities and reanalyzed them utilizing varimax rotations of the principal-factor matrices, following with additional rotations using the Schmid-Lieman procedure (Schmid & Lieman, 1957), which further orthogonalized the factors for a more parsimonious interpretation of the resulting factor structure.

Carroll concluded that a three-tiered model best fit the data. This model later became known as the 'three-stratum model.' The model stratified abilities according to breadth. The most general ability resides at the apex of the model at stratum III and is referred to as a general intellectual factor or *g*. The next level (stratum II) includes broad abilities (which are subsumed under the stratum III general factor) such as fluid intelligence (Gf), crystallized intelligence (Gc), general memory and learning (Gy), broad visual perception (Gv), broad auditory perception (Ga), broad retrieval ability (Gr), broad cognitive speediness (Gs), and reaction time/decision speed (Gt). Below Stratum II are over 70 narrow cognitive abilities which are organized according to their loadings on the various Stratum II broad factors.

Carroll's three-stratum model was widely embraced by the scientific community and represented a major paradigm shift in the study of cognitive abilities. The most significant contribution of the model is that it has provided the field of differential psychology with a standardized taxonomy to categorize and describe individual cognitive tasks. Despite its relative youth, Carroll's work is considered by many to be one of the greatest accomplishments in all of applied psychology. As Burns (1994) stated, "it is simply the finest work of research and scholarship I have read and is destined to be the classic study and reference work of human abilities for decades to come" (p. 35). In the nearly 20 years since its publication, Carroll's work has yet to be seriously challenged.

Consolidation and Refinement of the Gf-Gc Model

Despite the widespread acceptance of Carroll's theory within the academic community, several issues remained in coming to terms with where the theory fits within the broader three-stratum model. Despite congruence on many of the second-order factors, Carroll's model included a general factor whereas Cattell and Horn's models did not. This difference was a major source of contention for Horn, who refused to accept the validity evidence presented by Carroll for the general ability factor. As Horn and Noll warned, "the problem for the theory of general intelligences is that the factors are not the same from one study to another...The factors represent different mixture measures, not one general intelligence" (1997, p.68). Despite his reservations on *g*, Horn agreed to integrate the Gf-Gc model with Carroll's theory to form the broader CHC model (Newton & McGrew, 2010).

Since the consolidation of CHC, several empirical studies have tested Horn's theory regarding the exchangeability of *g* scores across studies. Compiling data from previous studies (*N* = 1,063), Floyd, Clark, and Shadish (2008) found that, despite the fact that differences in performance across tests may be observed for an individual, factor analysis indicated that the various general factors all loaded on the same latent variable construct. Additionally, Floyd and colleagues (2009) used generalizability theory to test the reliability of general factor loadings across various clinical samples (*N* = 1,409). They found that the general-factor loadings were moderately to strongly dependable across all of the samples examined, though the study was limited to analysis of only a single test battery. In a replication design using multiple cognitive tests administered to 433 individuals as part of the Minnesota Study of Twins Reared Apart (MISTRA), Major, Johnson, and Bouchard (2011) found that the general factor scores across tests loaded on the same latent *g* variable. Contrary to Horn's hypothesis, it appears that current intelligence tests all load on the same latent psychometric factor.

A remaining issue, and one that has resulted in the most research activity over the last decade, has been the question of how many broad factors should be represented at stratum II of the CHC model. In one of his last publications, Carroll (2003) concluded that there were data to clearly support 10 broad ability factors. For the better part of the last decade, there has been general consensus within the scientific community regarding most of these factors. Interestingly, in response to Woodcock's (1990) call for the inclusion of a reading and writing factor (Grw) in addition to Gq to form a more generalized acquired knowledge composite, Carroll did not embrace the inclusion of Gq at stratum II of his model. Instead, he classified quantitative reasoning as a narrow ability subsumed under Gf, which is in contrast with alternative models that place Gq at the stratum II level (e.g., McGrew & Flanagan, 1998). Carroll considered quantitative ability to be "an inexact, unanalyzed popular concept that has no scientific meaning unless it is referred to the structure of abilities that compose it" (1993, p. 627).

Since consolidation, McGrew's (2005) classifications of CHC abilities have become the standard framework for discussing CHC theory in the empirical literature. In a recent review, Schneider and McGrew (2012) provided updated definitions of the CHC stratum II abilities. The most commonly described broad abilities on current intelligence tests, adapted from Schenider and McGrew, and are provided below:

- *Fluid Reasoning (Gf)*: The ability to utilize both deductive and inductive thinking in order to solve novel problems. Fluid reasoning requires problem solving which cannot be completed by relying on previously learned schemas. Narrow abilities include induction, sequential reasoning, and quantitative reasoning.
- *Short-Term Memory (Gsm)*: The ability to encode, process, and manipulate information that is immediately available. Narrow abilities include memory span and working memory capacity.
- *Long-Term Storage and Retrieval (Glr)*: Defined as the ability to store, consolidate, and retrieve information over a period of time that extends beyond immediate awareness. Narrow abilities include learning efficiency, associative memory, and retrieval fluency.
- *Processing Speed (Gs*): The ability to perform elementary cognitive tasks efficiently. Narrow abilities include perceptual speed, psychomotor speed, and number facility.
- *Crystallized Ability (Gc)*: The depth and breadth of cultural knowledge that is largely the result of exposure to school-based learning tasks. Narrow abilities include verbal information, language development, and lexical knowledge.

- *Visual Processing (Gv)*: The ability to perceive, analyze, synthesize, and think with visual patterns, including the ability to store and recall visual representations. Narrow abilities include visualization, visual closure, and visual memory.
- *Auditory Processing (Ga)*: The ability to analyze, synthesize, and discriminate auditory stimuli, including the ability to process and discriminate speech sounds that may be presented under distorted conditions. Narrow abilities include phonetic coding and sound discrimination.

A graphic representation of the CHC stratum II abilities and additional hypothesized broad abilities is provided in Figure 1 (p. 215). It should be noted that these seven broad CHC abilities are the predominant focus of most of the empirical research on human cognitive abilities. Although recent work has demonstrated that some of the broad abilities are much more complex than previously thought (e.g., McGrew & Evans, 2004), such discussions are beyond the scope of this review. Nevertheless, it is important for practitioners to remember that the goal of contemporary CHC research is to continually refine the model so that it evolves into a more accurate summary of human cognitive abilities (Schneider & McGrew, 2012).

Impact of CHC Theory on Contemporary Test Development

In the decade plus that has elapsed since the consolidation of the Carroll and Cattell-Horn models, CHC theory has had a visible impact on the development of new and revised individually administered intelligence tests (Keith & Reynolds, 2010). According to McGrew and Schneider, “CHC theory has attained the status as the consensus psychometric model of the structure of human cognitive abilities” (2012, p. 109). To better understand this influence, the representation of CHC broad abilities on current intelligence test batteries will be reviewed next, along with the empirical evidence for CHC interpretations of those tests.

Wechsler scales. Although CHC theory is referenced within the manuals of the Wechsler Intelligence Scale for Children-Fourth Edition (WISC-IV; Wechsler, 2003) and Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV; Wechsler, 2008), the tests are not based explicitly on CHC theory. Both levels of the Wechsler scales purport to measure second-order indexes such as verbal comprehension (VCI), perceptual reasoning (PRI), working memory (WMI), and processing speed (PSI), both scales provide a third-order full-scale score. Whereas independent confirmatory factor analysis (CFA) supported a five factor CHC interpretation of the WISC-IV (Keith, Fine, Taub, Reynolds, & Kranzler, 2006), such an interpretation required splitting the PRI into separate fluid reasoning and visual processing indexes as well as the inclusion of several supplemental subtests to better complete the model. Interestingly, the resulting CHC model demonstrated better fit statistics when compared against the conventional Wechsler interpretive framework. However, exploratory hierarchical factor analyses on both the WISC-IV (Watkins, 2006) and the WAIS-IV (Canivez & Watkins, 2010) provided support for the Wechsler four factor solution. One of the major issues with interpreting the Wechsler scales from a CHC perspective is the fact that the second-order factor indexes are mixtures of multiple broad abilities. For example, the PRI contains a mixture of visual processing and fluid reasoning abilities. Thus, practitioners who interpret the factor solely as a measure of fluid reasoning ignore a host of construct-irrelevant variance that is hypothesized to be contained within the factor.

Differential Ability Scales. The DAS-II is a revision of a test that was derived from an eclectic mix of intelligence theory. Although the purpose of the DAS-II is to provide a robust estimate of overall cognitive ability through the selection of subtests which have the highest *g* loadings, the second edition makes reference to a CHC framework for interpreting subtests and factor scores despite largely retaining the structure of the first iteration of the scale. CHC broad abilities measured by the DAS-II include fluid reasoning, crystallized ability, visual processing, short-term memory, and processing speed. A third-stratum IQ composite is also provided. An independent confirmatory factor analysis (CFA) indicated that that a CHC framework fit best with the DAS-II factor structure and was invariant across age groups (Keith, Low, Reynolds, Patel, & Ridley, 2010).

Stanford-Binet. The Stanford-Binet Intelligence Scales-Fifth Edition (SB-V; Roid, 2003) is organized according to a broad ability structure of five factors composed of measures of fluid reasoning, crystallized ability, quantitative reasoning, visual processing, and short-term memory, which are all subsumed under a hierarchical *g*-factor score. Although CFA evidence provided in the test manual generally supported the five factor CHC solution, significant questions have been raised regarding the structural validity of the SB-V. Two independent exploratory factor analyses (EFA) using the SB-V standardization data supported only the presence of strong general factor (Canivez, 2008; DiStefano & Dombrowski, 2006). Although it should be noted that such structural validity issues have been noted for the Stanford-Binet scale since it was revised to a multi-factor ability scale for the first time with the publication of the previous edition in 1986. Keith and Reynolds (2010) speculated that the fact that it is the only major intelligence test that assesses all of its factors simultaneously through both verbal and visual formats makes the instrument too complex to model using conventional factor analytic methods.

Kaufman Assessment Battery for Children. Although the original Kaufman battery was designed to assess components associated with the Lurian model of mental processing (Luria, 1966), the Kaufman Assessment Battery for Children-Second Edition (KABC-II; Kaufman & Kaufman, 2004a) provides users with the option of interpreting from either the Lurian model or CHC. When interpreted from a CHC perspective the battery yields fluid reasoning, crystallized ability, long-term retrieval, short-term memory, and visual processing broad ability indexes. It also provides a third-order general ability index. Although validity evidence provided in the test manual supported the CHC model for ages 6 to 18, an independent CFA study found that from ages four to five, the fluid reasoning and visual processing indexes could be combined to form a joint factor (Morgan, Rothlisberg, McIntosh, & Hunt, 2009). Interestingly, no validity evidence was provided in the test manual to support interpretation from the alternative Lurian neuropsychological perspective.

Woodcock-Johnson. Prior to the development of CHC, the Woodcock-Johnson Psychoeducational Battery-Revised (WJ-R; Woodcock & Johnson, 1989) was the only major intelligence test to reference Gf-Gc theory. The WJ-R was designed to assess fluid reasoning, crystallized ability, long-term retrieval, short-term memory, visual processing, auditory processing, processing speed, and quantitative reasoning. The most recent iteration of the WJ cognitive assessment series, the Woodcock-Johnson III Tests of Cognitive Abilities (WJ-COG; Woodcock, McGrew, & Mather, 2001c) is the only current test based exclusively on CHC theory, and it is the only test purported to measure all eight of the CHC broad abilities specified within the latest iterations of the model. The WJ-COG was redesigned to better measure stratum I narrow abilities and provides two narrow measures for each broad ability composite. The WJ-COG also includes a general factor (*g*) at its apex. Since its publication in 2001, the WJ-COG has been utilized as a major research tool in the expansion and refinement of the CHC model. The proposed structure of the WJ-COG has been supported by CFA evidence (McGrew & Woodcock, 2001) and has been shown to be invariant across age groups (Taub & McGrew, 2004). Although a recent higher-order exploratory analysis conducted by Dombrowski (2013), using the correlation matrices for several school-age subgroups, did not support the CHC factor structure reported in the WJ-COG test manual. Principally, it was found that only the processing speed and long-term retrieval factors were well supported after the extraction of a strong general factor. Based upon these results, Dombrowski did not advise interpreting beyond the general factor until additional evidence supporting the viability of the proposed seven factor model was accumulated. Later, Dombrowski and Watkins (2013) utilized EFA procedures on correlational matrices that included all 42 subtests from the WJ-III battery and found evidence for additional broad factors, although the full CHC model structure for the WJ-COG was not supported. Nevertheless, since its publication, the WJ-COG has become one of the most popular intelligence test batteries administered in clinical practice by school psychologists. To date, it remains the only major intelligence battery that is purported to assess all hypothesized CHC broad factors without being confounded by alternative interpretive frameworks (e.g., the Luria model).

Summary

For most of the 20th century, authors in multiple fields of applied psychology have worked to develop and refine a comprehensive model of intellectual functioning and broad cognitive abilities. Those efforts have culminated in the development of the comprehensive CHC model. At present, many intelligence researchers accept the CHC model as a valid psychometric representation of the structure of human cognitive abilities. Accordingly, the CHC model has also had a dramatic impact on the development and refinement of current intelligence tests batteries. Most test batteries now reference CHC to varying degrees in their factor structures and/or interpretative frameworks. In fact, CHC theory is now being proffered as a useful architecture for the identification of specific learning disabilities in children and adolescents (see Flanagan, Alfonso, & Mascolo, 2011).

Despite the widespread representation of CHC within the cognitive testing landscape, some caveats should be noted. First, the WJ-COG is the only test battery founded exclusively on CHC theory and remains the only contemporary test that assesses all of the eight known broad ability factors. A survey of remaining tests with contrasting CHC influence revealed disproportionate representation across broad factors. Whereas coverage is widespread on fluid reasoning, crystallized ability, short-term memory, and visual processing, auditory processing and long-term retrieval are underrepresented within existing cognitive measures. Another interesting observation is the fact that many of the contemporary tests which are purported to incorporate new elements of CHC within their factor structure (e.g., Wechsler scales) retain most if not all of the factor structures and subtest compositions from previous editions. In some cases this results in the reporting of more factors without any discernible increases in content within the test structure. This finding has led some researchers (e.g., Dombrowski, 2013; Frazier & Youngstrom, 2007) to conclude that many contemporary intelligence tests, including those based upon CHC theory, may be over-factored. Despite these discrepancies, the results of a variety of psychometric investigations provide some support for the structural validity of the CHC model in contemporary intelligence tests within the empirical literature.

History of Intelligence Test Interpretation

As previous sections of this review attest, the historical development of intelligence theory and as the tests that have been proposed to measure the construct itself have been controversial. Naturally, such debates have impacted discussions regarding how practitioners should interpret the data obtained from intelligence test measures. Because these measures are utilized within educational settings to make decisions about a number of socially important outcomes (Daniel, 1997), the debate regarding how they should be interpreted best has been the focus of much speculation by intelligence test scholars for most of the last century.

As intelligence tests have become more sophisticated over time, the methods that have been proposed to interpret them have become increasingly more complex. Proponents of such approaches have argued that recent developments have made inferences derived from observations of test performance more valid and defensible, while critics challenge that the new methods are in fact nothing more than recycled versions of previously failed practices. In an elegant discussion, Kamphaus, Winsor, Rowe, and Kim (2005) presented a model to suggest that the history of intelligence test interpretation can be divided into four waves or phases. A new wave emerges as a result of gaps identified within the existing knowledge base, prompting a paradigm shift. Each wave and its impact on the practice of school psychology are discussed next in more detail.

First Wave (1905-1944)

The first wave of intelligence test interpretation lasted approximately from the birth of the first intelligence test until the middle of the 20th century, when the use of factor analytic methods to interpret modern intelligence tests became more pronounced. Interpretation during the initial period was driven by an effort to estimate an individual's general level of intelligence, predominately through the use of a full-scale score or IQ score. Early versions of IQ tests (e.g., the Binet-Simon scale) were criterion-referenced tests whereby a person's raw score was referenced to a preconceived cutoff that was thought to demarcate the expected level of cognitive functioning for an individual at that age. Interpretation was generally limited to estimations of a person's level of skill based upon item analysis and qualitative assessment of test performance.

With the introduction of standardization procedures, examiners began to be able to make inferences relative to an individual's performance compared to other individuals who took that same test under the same conditions (Pintner, 1923). Performance was then referenced to some kind of arbitrary classification hierarchy. Wechsler (1944) introduced a popular system that is still used by many clinicians today that attempted to tie categories (i.e., delayed, average, superior) to statistical frequencies found within the normal or Gaussian distribution.

A confluence of factors led to the demise of the first wave within clinical psychology. First, the emergence of multi-factor models of intelligence led many scholars to question whether cognitive ability could be reduced to a summary score. Gottfredson (2005b) argued that public unease with summary cognitive scores was fueled by the historic association between IQ and the eugenics movement, which may explain why theories which deemphasize IQ are popular among many clinicians. Subsequent revisions of intelligence tests followed suit and began providing additional scores and factors (i.e., verbal, non-verbal) which were previously unavailable for users to interpret. These factors, together with the publication of Rapaport, Gil, and Schafer's work (1945), elicited a new era of clinical profile analysis.

Second Wave (1945-1958)

The cornerstone of the second wave was a transition away from the general intelligence construct in favor of an analysis of the shape of a person's profile of subtest scores. The belief was that individual variations in cognitive test performance served as evidence for the presence of a variety of clinical disorders. The Rapaport et al. (1945) system of analysis included five methods of interpretation: a) survey of individual item responses, b) analysis of within subtest item responses, c) generating hypotheses from subtest profiles, d) analysis of potential discrepancies between verbal and performance skills, and e) contrasting obtained scores with other sources of data. The method provided clinicians with a step-by-step process for analyzing several levels of intelligence test scores and was the first system to advocate for the analysis of intra-individual strengths and weaknesses.

Although the second wave provided more sophisticated methods for analyzing intelligence test data, normative guidelines were lacking for evaluating the validity of such interpretations. Validation of such methods largely was limited to publication of individual case studies within clinical journals in which inferences were made between the obtained score performance of an individual and their documented clinical condition (e.g., schizophrenia). The lasting contribution of the second wave was the reification among clinicians and the lay public of the diagnostic importance of score differences. Such approaches also invited and validated hypothesis generation by the individual clinician, a practice which soon flourished with the rise of multi-factor ability tests and the rise of factor analysis.

Third Wave (1959-1996)

Commensurate with the proliferation of computer and statistical evaluation technologies, researchers soon began to investigate the psychometric properties of intelligence scales using factor analytic methods. In a study of the original WISC, Cohen (1959) provided evidence for the validity of a three factor interpretation of WISC score data (full-scale, performance, and verbal). The three-factor structure became the standard for interpreting the WISC as well as many other intelligence tests for many years to come. Interestingly, several additional factors were tested which were not retained in the three-factor model, one of which (freedom from distractibility) was included in a later version of the test. This study crystallized the importance of test interpretation that primarily was based on measurement science and not on clinical intuition, and ultimately led to the era of psychometric profile analysis.

Profile analysis, also known as ipsative test interpretation, is the process of comparing performance across individual subtests or indexes relative to the mean performance across all tasks. Scores below the mean are interpreted as a weakness whereas scores above the mean are inferred to be cognitive strengths. Although the concept of ipsative analysis was first introduced by Cattell in 1944, Davis (1959) was the first to develop a working formula for determining inter-scale strengths and weaknesses. The Davis formula utilized mean subtest performance across the entire Wechsler battery as the basis for comparison.

Alan S. Kaufman later wrote several influential books arguing for a synthesis of clinical and psychometric interpretation of cognitive test score data which he referred to as ‘intelligent testing.’ The Kaufman (1994) approach is a hierarchical multi-step method by which practitioners are encouraged to utilize all scores provided on an intelligence test to derive hypotheses about individual cognitive performance. The steps in the Kaufman approach can be summarized as follows: 1) determine the best way to summarize overall ability, 2) determine if observed index differences are significant, 3) determine if individual factor scores are unitary and interpretable, 4) interpret individual subtest strengths and weaknesses, and 5) generate hypothesis about observed index or subtest profiles.

Kaufman (1994) likened clinicians to detectives searching for clues and sifting through evidence. Whereas intelligence and psychometric theory are important, Kaufman placed particular emphasis on clinical acumen: “With experience, a well-trained examiner will be able to shift from one approach to another to find the best explanations for the observed fluctuations in a child’s profile and to infer cause-effect relationships between behaviors and test performance" (p. 23). Although the intelligent testing approached has been revised several times in response to critics, the core tenet of generating hypotheses from an integration of both qualitative and quantitative test data has remained constant. Despite its critics, the influence of the Kaufman approach on contemporary test interpretation cannot be overstated as it became "for many, the *sina qua non* of interpretation systems" (Fletcher-Janzen, 2009, p. 20).

Despite the promise of advances in psychometric profiling, research conducted over the last two decades has failed to validate the clinical usefulness of such approaches. Hypotheses about individual cognitive functioning inextricably are linked to the estimates of the validity and diagnostic utility of the individual scores from which such inferences are derived. Conventional guidelines (e.g., Nunally & Bernstein, 1994) for examining the psychometric properties of intelligence test scores within the field of school psychology are well established; it is within this psychometric framework that various test interpretations are evaluated within this review.

Criticisms of Interpretation at the Subtest Level

The psychometric deficiency of subtest level interpretation was first demonstrated by Cohen (1959). In a study of the standardization sample of the WISC ($N = 2{,}200$), Cohen examined variance partitions for each of the WISC subtests. Briefly, intelligence test factors and subtests are all composed of common, specific, and error variance. Common variance or communality refers to the portion of test score variance that is shared with all other factors or tests. On intelligence tests, common variance is most often attributed to the *g*-factor. Unique variance or specificity denotes the amount of variance that is unique to that test and is interpreted to represent the degree to which a test measures various broad or narrow cognitive abilities. Finally, error variance is a byproduct of the reliability or accuracy of the test measure and is not interpretable. In the Cohen study, few of the WISC subtests demonstrated specificity estimates higher than 30%. In some cases specificity estimates did not exceed the error variance in individual tests.

Ultimately, interpretation of subtest scores rests on the assumption that such scores are relatively stable. Canivez and Watkins (1998) examined the temporal stability of subtest scores on the WISC-III for successive special education evaluations conducted over a three year span ($N = 667$). They found that stability coefficients for individual subtests ranged from .55 to .75. Results such as these have been well replicated within the empirical literature (e.g., Livingston, Jennings, Reynolds, & Gray 2003; Watkins & Canivez, 2004).

Although the interpretation of subtest scores in isolation is not very common, interpretive systems such as Kaufman's which advocate elaborate intra-individual subtest comparisons are extremely popular amongst clinicians (Pfeiffer, Reddy, Kletzel, Schmelzer, & Boyer, 2000). Such approaches are termed 'idiographic' because strengths and weaknesses are relative to the individual and not based upon any kind of normative or base rate support. Thus it is not possible to determine if such scores represent a deficit when compared to other individuals at that age within the broader population.

In a review of existing studies of ipsative interpretive practices, McDermott, Fantuzzo, and Glutting (1990) found that the temporal stability for ipsative strengths and weaknesses on the WISC-R was insufficient for decision-making. The major problem the authors noted with ipsative analysis was that such methods rely on the calculation of difference scores which are even less reliable than the tests from which they are derived. In sum, McDermott and colleagues encouraged practitioners to “just say no” to subtest analysis.

McDermott, Fantuzzo, Glutting, Watkins, and Baggaley (1992) later demonstrated problems with using ipsative performance as a diagnostic methodology. They conducted a hierarchical regression analysis using WISC-R standardization data (*N* = 2,200), to determine if ipsative scores predicted achievement better than the normative scaled scores of the individual subtests. It was found that normative scores predicted 27% of achievement variance whereas ipsative scores predicted only 9% of the variance on the same dependent achievement variables. The authors concluded that once subtest scores are “ipsatized” all common variance is removed, presenting a significant confound for the interpretation of such scores.

Macmann and Burnett (1997) conducted a comprehensive analysis of the reliability of diagnostic decisions derived from the Kaufman method using simulation data on the psychometric parameters of the WISC-III. The authors found that the reliability of inferences made from both normative and ipsative analysis of subtest level performance rarely exceeded chance levels. Although the use of simulation methods to assess the psychometric properties of intelligence tests has been criticized by many test authors, Macmann and Burnett argued that such designs should be thought of as “best case scenarios” as many additional sources of error are not accounted for.

Despite the evidence calling into question the validity of subtest level interpretations of intelligence test batteries, such practices continue to be encouraged and described in many professional textbooks and manuals (e.g., Groth-Marnat, 2009). The issue is further obfuscated by inconsistent instruction provided in additional sources such as the popular "*Essentials of Psychological Assessment"* series published by Wiley. In a volume of that series, Flanagan and Kaufman (2004) encouraged clinicians to disavow ipsative methods when interpreting contemporary measures such as the WISC-IV. Curiously, in a subsequent text on test interpretation, adherents of the intelligent testing approach continued to extol the virtues of clinical subtest analysis: "Examiners must be detectives, actively attacking subtest profiles in systematic fashion. They need to group subtests in new ways to best explain each individual's subtest-to-subtest fluctuations" (Kaufman & Lichtenberger, 2006, p. 453). Such contradictions are commonly found when surveying recent work by Kaufman and colleagues.

Fourth Wave (1997-Present)

In the mid-1990s cognitive test interpretation was at a crossroads. The failure to validate previous clinical methodologies (e.g., subtest analysis) raised questions about whether intelligence tests were worth their cost and time to administer. A solution soon emerged as a new generation of tests began to report more complex factor structures, commensurate with the emergence of more sophisticated theories of intelligence. A fourth wave soon emerged at the turn of the 20th century. Because current ability measures provide multiple factor and index scores, test manuals began to include procedures for engaging in many of the same evaluative procedures that were previously utilized for subtest level analysis. Such methods were thought to have more clinical validity due to the fact that factor level scores do not suffer from the stability issues identified with subtests as a result of the Spearman-Brown prophecy (Traub, 1991). The assumption of such approaches is that significant differences observed among lower-level scores for an individual are indicative of cognitive strengths or weaknesses, in some cases calling into question the validity of the full-scale score. According to Weiss et al. (2006), large discrepancies “point to the need to shift test interpretation to the index score level where the most clinically relevant information is more likely to be found” (p. 140). By the turn of the century, factor level interpretation strategies were being taught in most school psychology training programs (Alfonso, Oakland, LaRocca, & Spanakos, 2000) and were preferred by most practitioners (Pfeiffer et al., 2000).

The fourth wave is characterized by a continued de-emphasis on the validity of the general factor and a focus on the integration of contemporary intelligence and neuropsychological theory within test development and interpretation. Whereas previous interpretive approaches advocated for atheoretical psychometric interpretations of subtests and some factors scores, contemporary approaches encourage practitioners to interpret intelligence tests using theory to guide hypotheses about how cognitive abilities relate and work together within a comprehensive framework. This goal is best accomplished by engaging in more selective forms of testing at the stratum II broad ability level. It is believed that advances in theory (e.g., CHC) provide users with more valid alternatives when interpreting contemporary intellectual tests. In the last decade, several interpretive frameworks which utilize tenets of CHC and other neurocognitive theory as their foundation have emerged which broadly can be characterized as ‘cross-battery’ in nature. That is, they encourage school psychologists and other clinical professionals to administer and interpret information from multiple batteries to develop a comprehensive profile of an individual’s cognitive strengths and weaknesses.

The Cross-Battery Assessment (XBA) method was first articulated on a theoretical level by Flanagan and McGrew (1997) and was later operationalized as a comprehensive interpretive method. XBA utilizes the CHC architecture to classify all factor scores on contemporary intelligence tests according to the broad abilities they are hypothesized to represent. Such taxonomy allows for practitioners to utilize CHC as an interpretive tool across batteries and is necessary due to the fact that only the KABC-II and WJ-COG exclusively utilize CHC terminology to describe second-order factors. According to Flanagan, Alfonso, and Ortiz (2012), the XBA approach "allows practitioners to focus on measurement of the cognitive constructs and neurodevelopmental functions that are most germane to referral concerns" (p. 459).

Although some aspects of the XBA method have been criticized on psychometric grounds (see Schneider & McGrew, 2011) the core tenets of selective broad ability testing and utilizing CHC as an interpretive taxonomy for evaluating cognitive processing profiles has been embraced in several models (e.g., Flanagan, Alfonso, & Mascolo, 2011) which recently have been proposed to replace the discrepancy model for identification of specific learning disability. Such models which emphasize the interpretation of broad ability strengths and weaknesses are referred to as *profile of strength and weaknesses* or PSW. PSW approaches to SLD identification currently are in the process of being developed by regulatory authorities in California (Special Education Local Plan Area Administrators of California, 2010) and Oregon (Hanson, Sharman, & Ezparza-Brown, 2009) thus, establishing the validity for interpretation intelligence tests at the factor level has broad policy implications for the field of school psychology and beyond.

The popularity of such interpretive frameworks has been buoyed by research that has demonstrated links between various broad CHC abilities and specific areas of academic functioning. Early research relationships between broad abilities and achievement took place toward the end of the third wave with the publication of the WJ-R. Multiple regression studies of the WJ-R revealed that various broad abilities consistently accounted for 50% to 70% of reading (McGrew, 1993), writing (McGrew & Knopik, 1993), and math (McGrew & Hessler, 1995) variance on standardized tests.

More recently, Evans, Floyd, McGrew, and Leforgee (2002) examined relationships between CHC factors on the WJ-COG and reading for school-aged children and adolescents (*N* = 7,641). Regression analysis revealed that crystallized ability, short-term memory, auditory processing, long-term retrieval, and processing speed had moderate to strong relationships with reading achievement throughout the school years, whereas no consistent relationships was found across the 14 age groups for fluid reasoning and visual processing. A stronger relationship was found between fluid reasoning and math achievement (Floyd, Evans, & McGrew, 2003), although once again the predictive utility of visual processing was found to be insignificant.

In a comprehensive synthesis of 19 studies, which included 134 separate analyses, McGrew and Wendling (2010) found significant support for links between broad cognitive abilities and several achievement areas. A major limitation of these earlier studies is that they did not control for the effects of the general factor. All of the regression studies cited above failed to include the full-scale score in the regression models. Failure to include such an important predictor variable in a regression equation constitutes a design flaw referred to as a specification error. According to Berry and

Feldman (1985), a specification error occurs when the “wrong model has been estimated” (p. 18). Model specification errors occur when an important variable is left out of the regression equation, and when a spurious variable is included within it. Regression equations that are poorly specified can result in prediction errors and models which leave a significant amount of dependent variable variance unaccounted for.

Given the well documented relationship between *g* and achievement outcomes, leaving *g* out of a predictive model assessing relationships between cognitive abilities and achievement is a particularly major specification error. Unfortunately, simply including *g* as a variable in a model along with subordinate cognitive factor scores introduces an additional confound. Due to the fact that factor scores are subordinate to the full-scale score in a hierarchical model, the relative merits of the residual scores can only be evaluated after first removing the predictive effects of common variance which is associated with the general factor. Such partitioning of predictor variables is important for establishing the incremental validity of independent variables.

To account for the effects of the general factor, researchers have incorporated structural equation modeling (SEM) into many predictive models. SEM is a multivariate method that combines multiple regression and factor analysis to assess relationships between multiple independent variables and one or more dependent variables. Although it is most often utilized to test theoretical relationships among variables, the results of SEM studies have implications for prediction (Thompson, 2000). Recently, researchers have utilized SEM methods to evaluate whether the general ability factor on intelligence tests has direct or indirect effects on achievement. In general, researchers using these techniques have found that *g* has strong but indirect effects on various achievement variables, in contrast to the broad factors which have more direct effects (Floyd, Keith, Taub, & McGrew, 2007; Vanderwood, McGrew, Flanagan, & Keith, 2001). Such studies are often proffered by proponents of factor level interpretation methods (e.g., Keith & Reynolds, 2010) as evidence that broad cognitive abilities must be considered when assessing more discrete areas of achievement. However, it is important to note that in SEM there is subjectivity inherent in model selection and testing (Lee & Hershberger, 1990). Ultimately, the validity of any SEM design is determined by how well it tests plausible alternative models. In the SEM studies cited above, it is difficult to discern exactly which competing models were tested and whether they were truly equivalent. Interestingly, investigators who have tested models which more explicitly specify direct interactions between *g* and various achievement variables have found less favorable results for the utility of factor level scores.

Oh, Glutting, Watkins, Youngstrom, and McDermott (2004) used SEM to analyze the predictive validity of WISC-III variables on reading and math achievement using the nationally standardized linking sample between the WISC-III/Wechsler Individual Achievement Test (WIAT; *N* = 1,116) and found that the effect sizes of relationships between the factor scores and achievement variables were small in comparison to the effect sizes obtained from the relationship between general ability and achievement. In a more recent investigation of the WISC-IV, Parkin and Beaujean (2012) tested several predictive models to determine which best accounted for relationships between cognitive variables (as measured by the WISC) and math achievement on the WISC-IV/WIAT-II linking sample (*N* = 550). Specifically, they tested a *g*-only model, a model with only the factor scores as predictors, and a hybrid model which utilized both the factor and full-scale scores. The authors concluded that *g* was the single best predictor of standardized math achievement.

An additional limitation of SEM is that it is a poor method for partitioning variance. To account for the relative contributions of subordinate scores, once must first control for the effects of common variance. SEM is able to do this at a tertiary level by assessing the indirect and direct effects of a factor or scale on an external criterion (e.g., achievement variable). However such models do not partition common variance away from the broad abilities or factors when assessing their effects, thus confounding the interpretation of SEM results. Taub, Keith, Floyd, and McGrew (2008) recognized this limitation and concluded that it was difficult to interpret the relative effects of some broad abilities on math scores because of the near perfect path loadings that were observed between those abilities and general intelligence.

Furthermore, the applied implications of SEM studies often are difficult to ascertain due to the fact that such methods assume that latent traits or constructs are measured without error. True score theory posits that all scores from intelligence test batteries contain measurement error (Crocker & Algina, 2008). Thus, practitioners who rely solely on the results of SEM investigations for identifying interpretation strategies for cognitive tests run the risk of over-interpreting spurious variables (Oh et al., 2004). As a remedy, Schneider (2013) recommended that practitioners estimate latent scores by transforming observed scores using established theoretical modeling parameters (e.g., CHC). Whether such procedures provide users with greater predictive ability in clinical contexts is not known. As Kline (2011) warned, the mere identification of a factor through latent variable modeling does not mean that the factor has clinical or diagnostic utility.

Summary

Since the publication of the first intelligence test by Binet and Simon in 1905 scholars have debated how such instruments should be interpreted. For most of the second half of the 20th century there has been a gradual shift away from strategies which emphasize interpretation of the full-scale or “IQ” score to methods which highlight identifying cognitive strengths and weaknesses via performance on subordinate subtest and factor measures. Such ‘sophisticated’ interpretive strategies assume that lower-level scores provide meaningful clinical information that cannot be accounted for by the full-scale score. Whereas subtest level analysis has been empirically challenged within the technical literature, contemporary interpretive methods (e.g., XBA, PSW) call for factor level interpretation for clinical decision making. Most of the evidence presented in favor of such arguments fails to account for applied aspects of psychometric validity (e.g., incremental validity) that are of interest to practitioners who utilize intelligence test data in clinical practice.

Incremental Validity of Intelligence Tests

The concept of incremental validity was first articulated by Sechrest (1963), who argued that psychological tests intended for use within clinical settings must provide users with meaningful information not yet accounted for by extant measures or available sources of data. For most of the second half of the 20th century incremental validity investigations were limited largely to investigating the legitimacy of various assessment methods within the field of clinical psychology. With the advent of the “intelligent testing" movement, such investigations were soon being conducted on intelligence tests by school psychology researchers.

In general, less convoluted explanations are favored when explaining observed phenomena within social science research, a scientific principle known as the law of parsimony which states "what can be explained by fewer principles is needlessly explained by more" (Jones, 1952, p. 620). When applied to intelligence tests, interpretation of the full-scale score is more parsimonious than interpretation at the factor-score or broad ability level. Thus, to interpret primarily at the factor level, practitioners should have a compelling reason for doing so. Whereas some researchers have questioned the value of such conservative scientific guidelines, Meehl (2002) argued that such guidelines are needed to protect the ability of researchers to falsify spurious hypotheses. Incremental validity of intelligence test variables is typically demonstrated through the use of hierarchical multiple regression procedures.

Hierarchical Multiple Regression

Hierarchical multiple regression (HMR) tests the proportion of dependent variable (DV) variance accounted for incrementally by all independent variables (IVs). That is, the additional variance explained by each IV is observed at the point at which it is entered into the regression equation. Variables are entered into the regression equation, one at time or in blocks, in a predetermined order by the researcher. The DV variance that is attributable to each predictor block is estimated by the squared multiple correlation coefficient (R^2, a statistical parameter of the population effect size f^2). R^2 is calculated by dividing the sum of squares due to regression by the sum of squares about the mean and can be interpreted as an effect size as a percentage of criterion variance accounted for. As an example, an R^2 coefficient of .69 is interpreted as indicating that 69% of total criterion variance explained. Guidelines for interpreting R^2 as an effect size are found in Cohen (1988); they are "small," .01; "medium," .09; and "large," .25. The critical coefficient in HMR analysis is the incremental squared multiple correlation coefficient (ΔR^2). The ΔR^2 represents the amount of variance that is explained by an IV in addition to the proportion of variance already explained by previous IVs in the regression equation. At present there are no conventional guidelines for interpreting the ΔR^2 coefficient.

Additionally, the incremental *F* ratio test (F_{inc}) can be used to assess whether a secondary block of variables provides for statistically significant increases above the variance predicted by a set of variables already in the prediction equation. The corresponding F_{inc} statistic is then interpreted according to whether it meets a predetermined level of significance (e.g., $p < .05$). The initial block of predictors can be tested using the traditional *F* ratio test.

Although regression coefficients (e.g., beta weights and unstandardized regression coefficients) can be used to estimate the effects attributable to individual variables, such statistics are rarely the point of interpretation in HMR studies due to the fact that they can be attenuated by threats to validity, whereas R^2 is robust to such threats (J. J. Glutting, personal communication, August 28, 2011). For example, when IVs are highly correlated (a threat to validity discussed in more detail below), the error terms for the corresponding beta weights become inflated. As a result, it is possible to conduct a significance test on a beta weight associated with a variable and fail to reject the null hypothesis (due to an inflated error term), whereas other regression statistics (e.g., R^2) can be interpreted as indicating no significant total effects within the same analysis.

HMR analysis depends on a number of statistical assumptions, including: an adequate number of cases for each of the IVs, absence of outliers among the variables, the use of orthogonal variables in the predictor equation, independence of prediction errors, use of variables that meet the assumptions for normality or homoscedasticity, and a linear relationship between predictor and criterion variables. It has been demonstrated that multiple regression generally is robust in the face of violations of assumptions (Pedhazur, 1997), except for measurement errors in the IV (e.g., use of unreliable measures) and specification errors (use of an inadequate predictor model in testing effects).

Previous Incremental Validity Studies Utilizing Regression Procedures

The first empirical test of incremental prediction on a contemporary intelligence test measure using HMR was conducted on the WISC-III. As previously stated, interpretation at the factor level is confounded by the fact that lower-order scores are saturated with common variance. In order to extract common variance from the WISC-III factors, Glutting, Youngstrom, Ward, Ward, and Hale (1997) utilized the HMR procedure to assess achievement outcomes with a clinical sample (n = 636) and a nationally stratified non-referred sample (n = 283). The full-scale score was entered into the regression equation first, followed by the four factor scores. Glutting et al. found that the full-scale score accounted for 29% to 55% of achievement variance across samples. Once extracted from the full-scale score, the factor scores accounted for an additional 5% to 15% of achievement variance. Although the results were statistically significant, effect size estimates (i.e., R^2 values ranging from .02 to .15) were fairly low utilizing Cohen's (1988) guidelines for interpretation. Youngstrom, Kogos, and Glutting (1999) later conducted a replication study on the DAS with a sample of 1,185 participants and found that the factor scales contributed only 2% to 4% additional predictive variance across achievement measures beyond the general factor. Interestingly, general factor variance estimates were substantially lower than those obtained in the WISC-III study by Glutting and colleagues, ranging from 15% to 35%.

The impact of significant inter-factor variability on the incremental validity of the DAS was subsequently examined by Kotz, Watkins, and McDermott (2008). Using the DAS national standardization sample of 1,200 children and adolescents, Kotz et al. utilized a matched case design whereby two groups were created so that they differed on the degree of broad index score discrepancy on the DAS, and matched on the full-scale score results. The results of an HMR analysis revealed no significant differences between groups in predicting achievement across several discrepancy levels. Similar results have been found by different researchers (see Canivez 2013b for a review) when examining inter-factor variability on several additional cognitive measures using a variety of samples and clinical conditions.

Watkins and Glutting (2000) later conducted an investigation to examine the incremental validity of profile status in predicting achievement on the WISC-III. Using data obtained from a mixed clinical sample of over 1,600 students, they created subgroups based upon factor profile scatter and shape. These categorical variables were then tested using HMR against profile elevation (as a proxy for general ability level). Across achievement variables, profile status provided 0% to 8% additional predictive variance beyond the general ability level.

HMR methods have only recently been utilized to assess the latest iterations of several popular intelligence tests. In a highly influential article Glutting et al. (2006) utilized a replication design on a WISC-IV/WIAT-II linked subsample (*N* = 498). They found that the full-scale score accounted for over 59% of the variance across reading and math outcomes, whereas the factor scores provided 0.3% to 1.8% additional predictive variance. The authors concluded that the results obtained from incremental validity studies using HMR are directly applicable to practitioners because the method assesses the utility of score data at the observed level.

The predictive validity of the WAIS-IV was assessed by Canivez (2013a) in a design that utilized a mixed sample of participants who were administered the WAIS-IV and WIAT-III (*n* = 93) as well those who were administered the WAIS-IV and newly revised WIAT-III (*n* = 59), although it is not clear from the study whether this was a clinical or standardization sample. Across various WIAT-II composites, the full-scale IQ score accounted for 42.7% to 76.7% of achievement variance, whereas the factor scores only accounted for 1.4% to 9.9% additional variance. On the WIAT-III, the full-scale IQ score accounted for 25.9% to 62.8% of achievement variance, and the factor scores contributed 5.1% to 11.7% additional prediction of achievement outcomes.

Canivez (2011) was the first to assess the incremental validity of a cognitive instrument founded upon an alternative theory of intellectual/neuropsychological functioning. Using standardization data (*N* = 1,600) obtained from the test author, Canivez applied a replication design to evaluate the predictive validity of the factor scores on the Cognitive Assessment System (Naglieri & Das, 1997). Across broad achievement scores on the WJ-R, the CAS full-scale composite score accounted for 37%

to 49% of the variance, whereas the factor scores contributed an additional 1% to 3% of predictive variance. These results indicate that regardless of theoretical orientation, the predictive validity of factor scores remains relatively low.

Despite the promises of sophisticated interpretive methods and the documented impact of Cattell-Horn-Carroll theory on the development of cognitive tests, there have been few empirical investigations of the incremental validity of CHC broad ability factors on current intelligence tests. In the only investigation using HMR to date, McGill and Busse (2012) examined the validity of the CHC factors on the KABC-II. Using standardization data ($N = 2{,}024$), they found that the full-scale composite accounted for 43% to 54% of the variance across achievement domains, whereas the CHC factors together provided an additional 1% to 4% of explained variance. Although the incremental prediction provided by the CHC factors was statistically significant, the effect size estimates were trivial (R^2 values ranging from .01 to .04).

It is important to note that not all incremental validity studies using HMR have demonstrated robust results for the supremacy of the general factor. For example, Nelson and Canivez (2012) found that the factor scores on the Reynolds Intellectual Assessment Scales (RIAS; Reynolds & Kamphaus, 2003) provided significant improvements (up to 70% of additional predictive variance) in the prediction of achievement measures beyond the full-scale score in a clinical sample of 521 adult participants. Interestingly, the full-scale score on the RIAS also accounted for much less achievement variance (9% to 30%) when compared to the amount of variance explained by composites from other intelligence tests. Similar results were observed in a study intended to replicate the results of previous investigations on the criterion validity of the WAIS-IV with a referred sample of 300 college students. The results of the HMR analyses by Nelson, Canivez, & Watkins (2013) indicated that the full-scale IQ score accounted for 13.9% to 49.1% of the variance across achievement tasks on the Woodcock-Johnson III Tests of Achievement (WJ-ACH) and Nelson-Denny Reading Tests, whereas the factor scores contributed an additional 4.4% to 29.9% of predictive power. It is worth noting that across the 10 achievement indicators, the WAIS-IV factor scores contributed an average of 74% of additional predictive variance, which was determined to be clinically significant. Such results raise the question of whether the results obtained from HMR studies using standardization data are invariant across more diverse samples.

Collinearity

Critics (e.g., Hale & Fiorello, 2004) have argued that the use of HMR to assess the incremental validity of intelligence tests is confounded by the fact that many of the factor scores on such measures are highly correlated with their corresponding full-scale composites, a statistical threat known as collinearity or multicollinearity. Collinearity occurs in regression analysis when one or more IVs are significantly correlated with each other, which impacts the standard error for the corresponding standardized regression coefficients, thereby increasing the chances of Type II error (i.e., the likelihood of rejecting a false null hypothesis; Gordan, 1968).

Collinearity is commonly assessed using one or more of several indicators available in most statistical programs. The most commonly used indicators are the *tolerance index*, *variance inflation factor* (VIF), and *condition index*. The tolerance index is a measure of the degree to which variables are independent from one another (Darlington, 1990). Tolerance values can range from 0 to 1, with lower values indicating overlap. VIF indicates the degree to which observed regression coefficients deviate from what would be expected from uncorrelated variables. A common interpretive heuristic is that collinearity is a concern when VIF is equal to or larger than 10. However Cohen, Cohen, West, and Aiken (2003) encouraged more conservative guidelines by lowering the VIF threshold to six or seven. The condition index is a measure of dependency of a particular variable on the others. An elevated condition index is associated with inflation in the standard error of the parameter estimate for a variable. Suggested criteria by Belsey (1991) for determining a condition index for a particular eigenvalue root is elevated are a condition value of 30 and above for a given dimension, coupled with variance proportions greater than .50 for at least two different variables.

In most of the studies in this review, the authors reported inflated collinearity outcomes which call into question the accuracy of the corresponding standardized regression coefficients (e.g., beta weights). However, there are no clear guidelines as to when collinearity becomes sufficient enough to invalidate the results obtained from regression studies (Pedhazur, 1997). The problem is compounded by the fact that there are no quick fixes if such collinearity is detected. One potential remedy (ridge regression) is so convoluted that most authors (e.g., Fox, 1991; Tabachnick & Fiddell, 2007) advise against its use. The most parsimonious solution, and the one most commonly recommended in contemporary statistical texts, is to simply eliminate the offending variables until collinearity is no longer an issue. If applied to incremental validity studies, this most likely would mean eliminating the factor scores from the regression equation, which in and of itself validates the null hypothesis for most of the studies that have been reviewed. Although it is possible to choose to eliminate the full-scale score in lieu of the factor scores, this most likely would leave an under-identified prediction model, resulting in a specification error.

It should be noted that the impact of collinearity on incremental validity studies is less of an issue than in other research designs due to the fact that researchers rarely interpret regression coefficients to assess the relative importance of IVs. Schneider (2008) argued that interpretation of the R^2 (squared multiple correlation) coefficient in the face of collinearity does not violate any statistical rules of thumb. This rationale is important because R^2 is the proportion of the dependent variable variance accounted for by optimally weighted IVs in an HMR design. Furthermore, Pedhazur (1982) argued that in HMR collinearity does not bias the ΔR^2 statistic, which is the primary focus of interpretation in incremental validity investigations because it reflects the additional variance explained after controlling for the effects of previously entered IVs. Thus, despite the criticism by Hale and colleagues, the continued use of HMR to assess incremental validity within the behavioral sciences appears to be appropriate (Berry, 1993; Cohen et al., 2003).

Regression Communality Analysis

Hale, Fiorello, Kavanagh, Hoeppner, and Gaither (2001) argued that the results obtained in HMR studies were artifacts of force entering the full-scale score into the regression equation first; by doing so, little variance was left over to implicate the factor scores. Using a sample ($N = 174$) of children classified with a learning disability, they conducted multiple studies using both HMR and regression communality analysis. They chose to enter the factor scores from the WISC-III jointly into the first block followed by the full-scale score. Interestingly, the factor scores accounted for 22% to 34% of the achievement variance, whereas the full-scale score failed to account for more than 1% of additional variance. As a remedy to the observed collinearity, Hale et al. encouraged the use of regression communality analysis for assessing the incremental validity of intelligence tests.

Communality analysis is a method of variance partitioning designed to identify the proportion of variance explained by the independent variable entered last in a regression equation (Pedhazur, 1997). The key difference between communality analysis and HMR is that in the latter, common variance partitioned from all of the factors entered together is used as a proxy for the full-scale score, whereas in HMR the full-scale score is used as a dependent variable within the regression equation itself. Communality procedures have been utilized to demonstrate the validity of interpretation at the factor level across test batteries and various clinical samples (Fiorello, Hale, McGrath, Ryan, & Quinn, 2001; Hale, Fiorello, Bertin, & Shermin, 2003).

The use of communality analysis as a method for assessing incremental validity is controversial. In 2007, a special issue of *Applied Neuropsychology* was commissioned by the journal editor to debate the use of such methods. Fiorello and colleagues (2007) submitted a paper replicating their earlier results on the WISC-IV with a mixed clinical sample of children diagnosed with learning disabilities (*n* = 128), ADHD (*n* = 71), and traumatic brain injury (*n* = 29). Subsequent commentaries (e.g., Dana & Dawes, 2007) were critical of the conclusions reached by the Fiorello research group and demonstrated through simulation data that the full-scale score remained robust despite fluctuations in index score unity.

Schneider (2008) later criticized the use of communality analysis for explanatory purposes, likening it to the use of an “Ouija Board”. In an analysis using simulation data, he demonstrated that the higher-order communalities that result from conducting communality analysis on factor scores from intelligence tests cannot be interpreted as measuring the same common variance as a full-scale composite score. This is because when factors used in communality analysis are imperfect indicators of *g*, the lower-order communality estimates leave a significant amount of *g* variance unaccounted for. By the time one gets to the higher-order communality most of the *g* variance is gone, leading to the erroneous conclusion that the lower-order communalities (as a proxy for unique variance) account for more achievement variance than the resulting full-scale score proxy communality.

In a rebuttal, Hale, Fiorello, Kavanagh, Holdnack, and Aloe (2007) continued to challenge the validity of HMR methods, citing the fact that factor scores account for relatively the same amount of total achievement variance when entered first in regression equations, which indicates that multicollinearity makes order of independent variable entry completely arbitrary. They wrote: “If these authors wish to cite these dubious ‘incremental validity’ papers, they should report the statistical fact that order of independent variable entry determines whether the FSIQ means everything…or nothing” (p. 41). However, model choice is not arbitrary and is guided by existing theory about how independent variables relate to one another (Pedhazur, 1997). Choosing to interpret intelligence tests at the factor level not only violates the scientific law of parsimony; it cannot be reconciled with existing intelligence theory (Oh et al., 2004). Despite admonitions, the use of inappropriate statistical procedures (e.g., regression communality analysis) for explanatory purposes in the assessment of incremental validity continues (see Elliott, Hale, Fiorello, Dorvil, & Moldovan, 2010).

Summary

In general, this review of the incremental validity research revealed that, across several instruments, the full-scale score predicted approximately 30% to 60% of achievement variance on standardized tests. The additional variance attributable to second-order factor scores tended to range from 0% to 15% when subjected to HMR analyses. Although more robust evidence for the predictive validity of factors has been found by researchers using alternative block entry methods and designs (e.g., communality analysis), the use of such methods has been criticized within the technical literature. To summarize, Kahana et al. (2002) concluded that interpretation at the factor level "does not significantly help in predicting achievement, even in specific content areas, and results in models that are more complex, confusing, and time consuming" (p. 91). However, most of the research that supports such a claim is based on assessments of the validity of the Wechsler scales and other similar tests (e.g., DAS) that are not modeled with CHC theory in mind. As an example, only DAS subtests with the highest *g* loadings contribute to form the full-scale composite (Elliott, 1990); thus, it would not be expected that subordinate scores would provide much incremental prediction once the high amounts of common variance associated with the general factor were extracted.

The conclusion that only the general factor is relevant in clinical practice has been rightfully criticized as being overly reductionist (Brody, 1997). Proponents of such an approach to test interpretation fail to account for the fact that factors other than *g* account for over half of standardized achievement variance (Detterman, 2002), although questions remain as to what those factors are and whether they provide diagnostic value on current tests of intelligence. Research on intelligence tests modeled to assess CHC theory has been less pronounced. The only study to assess such variables that could be located to date was an incremental validity investigation of the KABC-II conducted by McGill and Busse (2012). Although primary interpretation at the factor level was not supported, it should be noted that the KABC-II utilizes a dual theoretical structure and remains heavily influenced by its Lurian roots (i.e., simultaneous-successive processing).

Gaps Identified in the Literature

Several gaps are apparent in the existing incremental validity research base based on this survey of the literature. Principally, the majority of incremental validity research has been conducted on various iterations of the Wechsler scales. Generalizations from such research are problematic on methodological and theoretical grounds. From a methodological standpoint, many of the linking samples utilized were obtained from commercial standardization projects and ranged from 50 to 500 cases. Whereas such numbers are adequate for conducting HMR analyses, they provide for mediocre sample power (Cohen, 1988). For example, Freberg, Vandiver, Watkins, and Canivez (2008) found that the factor scores on the WISC-III failed to account for meaningful achievement variance beyond that already accounted for by the full-scale score across several conditions of inter-factor variability in an archived referred sample (*n* = 202). However, post-hoc power analysis indicated that the sample size precluded the researchers from detecting clinically significant effect sizes in several of their analyses. Unfortunately, threats to statistical conclusion validity rarely have been discussed in incremental validity research.

Furthermore, none of the reviewed studies accounted for the potential effects of age and/or school-level on the predictive validity of intelligence tests. Carroll (1993) concluded that the general factor was a more robust predictor of early achievement because of the predominant focus on foundational skills (e.g., early literacy components) within the curriculum. However, recent research (e.g., Swanson, Zhang, & Jerman, 2009) has provided some evidence for developmental differentiation in the relationships between broad factors and academic achievement as a result of curriculum changes that emphasize comprehension and content over basic skills as children progress through school.

Additionally, the way in which the full-scale composite on the Wechsler scales is constructed may attenuate the variance proportions attributed to the factor scores by inflating collinearity among the independent variables. A certain amount of redundancy can be expected on all intelligence tests when analyzing the relative contributions of various scores due to the fact that all full-scale composites are derived via a linear combination of the subordinate subtest scores. What potentially inflates collinearity in the Wechsler Scales (as well as all other tests with the exception of the WJ-COG) is the use of an arbitrary weighting procedure in which performance across subtests is averaged to create the full-scale composite. The WJ-COG is the only test for which subtests are weighted according to their relative *g* loadings, which is consistent with existing intelligence theory. However, an examination as to the effect of such a differential weighting scheme on the incremental validity of an intelligence test, such as the WJ-COG, has yet to be conducted.

Over-reliance on the Wechsler scales is problematic on a theoretical level due to the recent developments in differential psychology which have led to the rise of CHC theory. As previously stated, CHC theory is now embedded within the architecture of contemporary intelligence test development and interpretation. However, a comprehensive assessment of the incremental validity of an intelligence test measure founded solely upon CHC theory (e.g., WJ-COG) has yet to be conducted. This assessment is critical given the questions that have been raised regarding the viability of the proposed WJ-COG factor structure (Dombrowski, 2013).

Finally, the results obtained from several incremental validity studies using referred samples (e.g., Nelson & Canivez, 2012; Nelson, Canivez, & Watkins, 2013) indicate that there are potential conditions that are commonly encountered by practitioners in clinical practice (e.g., significant inter-factor variability) that may mitigate the ubiquity of the general factor in predicting achievement. However, no incremental validity studies have accounted for the potential effects of complex cognitive mediators such as SLODR. Therefore, an incremental validity investigation of the WJ-COG that utilizes an adequate sample size and accounts for several potential mediating clinical conditions has the potential to provide meaningful information to the existing school psychology literature base.

Purpose of the Current Study

The purpose of this study is four-fold. First, the incremental validity of the CHC broad factors in predicting various achievement outcomes after controlling for the effects of the general factor will be assessed using data obtained from the broader Woodcock-Johnson III (WJ-III; Woodcock, McGrew, & Mather, 2001a) assessment battery standardization sample. This examination will be completed to orthogonalize specific variance that is unique to the various CHC factors from the common variance attributable to *g*, which is necessary to examine predictive relationships between cognitive abilities and academic achievement. The second part of this study is designed to assess whether incremental validity is invariant across different levels of schooling. The third part of the study is designed to examine whether the predictive validity of the general factor is impacted by the presence of various levels of significant inter-factor variability. In other words, is the general factor impacted when significant differences are observed between subordinate variables from which it is composed? The fourth part of the study is designed to assess whether the phenomenon of SLODR has an impact on the predictive validity of the general factor. That is, does the full-scale score retain established levels of predictive validity across different groups which are designed to reflect the tail ends of the normal distribution? The later areas of the study are important in establishing the consistency of incremental validity results across various clinical conditions.

Importance of Current Study

The current research is important for reasons that have implications for cognitive theory and clinical practice. As was demonstrated in this chapter, many scholars and practitioners may be ambivalent about the nature and importance of the general factor as it relates to the assessment and treatment of learning difficulties. Recently proposed models of SLD identification emphasize primary interpretation at the broad factor level at the expense of the full-scale score, which is thought to reflect general intelligence. Such models rest on the assumption that broad factor scores account for significant amounts of achievement variance beyond the full-scale score. Incremental validity investigations assess this important aspect of the relationship between cognitive abilities and academic achievement.

To date, a comprehensive incremental validity investigation using HMR on an assessment battery founded solely on CHC theory has yet to be conducted. As previously discussed, the WJ-COG is the only current battery based on CHC theory. Interestingly, such research has yet to be conducted on any iteration of the Woodcock series. A design utilizing the WJ-COG is critical on a theoretical level due to the fact that the WJ-COG has been utilized as the primary instrument for validating many of the refinements to CHC theory in the empirical literature over the past decade. The results of this investigation potentially will provide information relevant to assessing the validity of various PSW assessment models and the clinical utility of the CHC model (as measured by the WJ-COG). This study potentially also will provide practitioners with important information that is relevant for determining more accurate methods for interpreting the WJ-COG in clinical practice. Additionally, this study is one of the first to examine incremental predictive relationships between cognitive abilities and achievement exclusively from a CHC-based perspective.

According to Carroll (1993), to establish the validity of CHC broad factors, it is necessary to study variation in task performance as a function of type of task and status on the various broad ability factors, while controlling for the effects of the third-order general factor. The current study is designed to meet this need and to provide information for evaluating the efficacy of WJ-COG with respect to the commentary for standard 1.12 in the joint testing standards which states: “When a test provides more than one score, the distinctiveness of the separate scores should be demonstrated,” (American Educational

Research Association, American Psychological Association, National Council on Measurement in Education, 1999).

Research Questions and Hypotheses

Based upon the current literature, this research study is designed to evaluate the incremental validity of the WJ-COG by exploring the following research questions:

1. Do the CHC broad ability factors on the WJ-COG provide *statistically* significant incremental prediction of achievement outcomes above and beyond the effects of the general factor?
2. Do the CHC broad ability factors on the WJ-COG provide *clinically* significant incremental prediction of achievement outcomes above and beyond the effects of the general factor?
3. Is the predictive validity of the CHC factors on the WJ-COG invariant across different levels of schooling?
4. Is the predictive validity of the general factor attenuated by significant levels of inter-factor variability on the WJ-COG?
5. Does Spearman's law of diminishing returns (SLODR) impact the predictive validity of the general factor on the WJ-III COG?
6. Does the use of a differential weighting scheme enhance the validity of the WJ-COG factor structure in predicting norm-referenced reading, math, and writing outcomes, when compared to estimates that have been obtained from other intelligence tests using similar methods of variance partitioning on commercial standardization samples?

Although this study is primarily exploratory in nature, several hypotheses are offered based upon a review of theory, and current and past research on topics related to this study.

Hypothesis 1: The CHC factors on the WJ-COG will account for *statistically* significant amounts of incremental achievement prediction on the WJ-ACH above and beyond the effects of the general factor. For the first part of this study it is expected that the CHC factors will provide statistically significant predictive variance beyond the general factor. During model testing, when the general factor was entered first and the factor scores were entered into a second block, the second block accounted for significant effects (i.e., $p < .05$) in all of the HMR studies that were cited earlier. Given the relatively large sample size being utilized in this study, it was expected that this outcome would be replicated here with the WJ-COG. Such a finding will provide evidence for the predictive validity of the CHC factors at the statistical level.

Hypothesis 2: The CHC factors on the WJ-COG will not account for *clinically* significant amounts of incremental achievement prediction on the WJ-ACH above and beyond the effects of the general factor. In contrast to the first research question, the criterion validity of the WJ-COG CHC factors will be assessed to determine their practical or "clinical" significance. Whereas statistical importance is evidenced by obtaining findings on tests of significance that exceed *a priori* chance levels, clinical significance is assessed through effect size estimates, confidence intervals, and measures of association (Stevens, 2009). Most of the HMR studies cited above have resulted in weak effect size estimates associated with factor scores on most contemporary intelligence tests (Canivez, 2013b). Although conflicting results have been found in several studies (e.g., Nelson & Canivez, 2012; Nelson, Canivez, & Watkins, 2013), such exceptions have been limited to investigations with clinical samples of adult participants. Because this study utilized a commercial standardization sample with school-age participants, it was expected that previous effect size estimates would be replicated on the WJ-COG.

Hypothesis 3: The predictive validity of the CHC factors will not be invariant across levels of schooling. Although contemporary intelligence texts often refer to the ubiquity of the general factor in predicting school achievement, questions have been raised as to how invariant predictive validity estimates are across age groups. As was previously noted, incremental validity studies conducted on adult populations (e.g., Nelson & Canivez, 2012; Nelson, Canivez, & Watkins, 2013) presented evidence of a weakened general factor and greater incremental validity at the broad ability level in predicting various achievement tasks. An implication of such studies is that broad cognitive abilities become more of a factor in predicting achievement as a result of the fact that curriculum becomes more complex as students become older (i.e., shifting from emphasis on basic skills to comprehension and idea production). As a result, it is anticipated that the predictive validity of the CHC factors will not be invariant across levels of schooling (e.g., primary versus secondary) as the broad factors become more developed and play a greater role in mediating achievement.

Hypothesis 4: The predictive validity of the general factor will not be attenuated by significant levels of inter-factor variability on the WJ-COG. Although Watkins and Glutting (2000) indicated that the general factor remains a robust predictor of achievement when the subordinate factors that compose it are significantly discrepant, Hale and Fiorello (2001) questioned whether *g* remains a viable construct in clinical practice given the emergence of recent cognitive and neuropsychological theories of intelligence (e.g., CHC) which support individual differences at the broad and narrow level of abilities. In the absence of any specific evidence of the impact of inter-factor variability on the predictive utility of the general factor, it was expected that the general factor would remain a robust predictor of achievement variance across several levels of significant inter-factor variability.

Hypothesis 5: Differential rates of predictive validity will be demonstrated by the CHC factors on the WJ-COG across groups classified by estimated general ability level. Reynolds (2013) provided evidence in support of the differentiation hypothesis known as SLODR and its potential impact on the ubiquity of the general factor. Namely, it has been demonstrated that positive manifold among cognitive tests decreases as general ability rises. Because *g* loadings are associated with common variance, it was expected that differential incremental prediction would be evidenced across general ability groups. More specifically, it was anticipated that the CHC factor scores would provide more robust incremental prediction in a group composed of individuals with advanced levels of general ability when compared to estimates derived from competing groups stratified according to average and below average levels of general intelligence.

Hypothesis 6: The predictive utility of the CHC factor structure on the WJ-COG will be consistent with model estimates that have been obtained from HMR analyses of other intelligence tests. Previous HMR analyses of other intelligence tests consistently have found that proposed test models have accounted for approximately 50% to 60% of total achievement variance, when using data obtained from commercial standardization samples. However, none of these studies utilized a differential weighting procedure for the full-scale composite. Such a procedure potentially may reduce the attenuating effect of some of the dependence between the factors and the resulting full-scale composite. Nevertheless, the use of such a procedure was not expected to have a significant impact on prediction when compared to estimates obtained from other intelligence tests. Recent factor analytic investigations (e.g., Dombrowski, 2013; Dombrowski & Watkins, 2013) indicated that the CHC factors on the WJ-COG are highly saturated with common variance attributable to the general factor.

Chapter III: Methods

Sample and Participant Selection

The participants for the current study were drawn from the national standardization sample of the Woodcock-Johnson III (WJ-III; Woodcock, McGrew, & Mather, 2001a). The total standardization sample for the WJ-III is nationally stratified and consists of 8,818 participants between the ages of 1 to 90 years. The WJ-III technical manual (McGrew & Woodcock, 2001) provides evidence that the total sample meets or exceeds established demographic targets from the 2000 U.S. census for region, sex, and race. On the basis of the information provided in the technical manual, the WJ-III norms meet established standards (e.g., American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999) for norm construction. The normative scores on the WJ-III were weighted according to the 2000 census estimates. In 2007, a normative update was completed on the WJ-III (WJ-III NU; Woodcock, McGrew, Schrank, & Mather, 2007) to recalibrate normative weights in the computer scoring program to reflect 2005 U.S. census estimates. It should be noted that variable weights were recalibrated using participant data from the original 2001 WJ-III standardization study, no additional sampling was conducted. All standard scores in the present study are derived from weights from the WJ-III NU revision.

Specifically, this study utilized a subsample from the total WJ-III standardization study. The subsample consisted of school-aged children and adolescents who were administered portions of both the Tests of Cognitive Abilities (WJ-COG) and Tests of Achievement (WJ-ACH; Woodcock, McGrew, & Mather, 2001b). Although the larger WJ-III normative sample includes a much wider age range of participants, this study was restricted to school-aged participants because these are the age ranges most often assessed by practicing school psychologists in clinical practice. The target sample excludes the typical age-range (e.g., 3 to 5) that is associated with preschool practice. The decision not to include those cases was made to avoid several measurement confounds that are unique to preschool psychoeducational assessment (see Nagle, 2007). Most germane to the current study is the implication that the CHC factor structure may not be invariant across age groups. In an analysis of the CHC factor structure proposed for the KABC-II, Kaufman and Kaufman (2004b) found that the fluid reasoning factor could not be distinguished from the visual processing factor in the final structural model for the 4 to 6 age group. In addition, the WJ-ACH does not provide cluster or index scores for several achievement variables (e.g., reading, math, writing) for ages 3 through 5.

The current study sample consists of 4,722 cases between the ages of 6 and 18 years. Table 1 (p. 183) presents the relative proportions across demographics for sex, race, ethnicity (Hispanic/Not-Hispanic), region, community type, and parent educational level for the sample along with comparable 2005 U.S. census estimates. Equivalence between a sample and national anchor norms is typically demonstrated by an adequate apportionment of cases across relevant demographic variables (e.g., 100 cases or more) as well as well as satisfying the equipercentile method (Salvia & Ysseldyke, 2007). According to Anastasi and Urbina (1997), the equipercentile method refers to utilizing the distribution of proportions for relevant demographic variables in the population as reference markers for evaluating the adequacy of a study sample. A survey of the variables included in Table 1 (p. 183) indicates that only the number of participants who identified themselves as American Indian (*n* = 95) and those who reported being home schooled (*n* = 52) are underrepresented within the sample. The distribution of cases across age and grade ranges is presented in Table 2 (p. 184). Both ranges are adequately represented with the average number of cases for the K through 12 gradients (*M* = 348.92) and age gradients (*M* = 363.23) all exceeding minimum thresholds (*n* > 100). Interestingly, grade identifiers were not available for 54 cases (1.1%) in the sample. Equivalence between the demographic proportions obtained in the sample and those that are estimated from the 2005 census data was assessed using the chi-square goodness of fit test using a modified formula for estimating proportions by Hopkins (1979). Chi-square reflects the degree to which observed proportions fit with expected or theoretically derived proportions for a given category (Glass & Hopkins, 1996). Statistically significant differences (i.e., $p < .05$) were observed between the sample and estimated population parameters in race, χ^2 (3) = 36.83, ethnicity, χ^2 (1) = 139.40, region, χ^2 (3) = 133.35, size of community, χ^2 (2) = 597.81, type of school, χ^2 (2) = 28.80, place of birth, χ^2 (1) = 5.19, father's educational level, χ^2 (2) = 12.60, and mother's educational level, χ^2 (2) = 29.70. No statistically significant differences were noted for gender representation within the sample, χ^2 (1) = 1.21. Due to the size of the sample, trivial differences between observed and expected proportions often resulted in significant chi-square tests. On a practical level, the average difference across all demographic variables was relatively small (*M* = 2.57), although it should be noted that thresholds for determining when observed differences should be interpreted as clinically significant have not been established within the literature. Therefore it is believed that the sample for the current study provides for adequate estimation of the established population parameters, thus the use of inferential statistical analyses is warranted.

Measurement Instrument

Woodcock-Johnson III. The WJ-III is an individually administered, norm-referenced assessment system for the measurement of overall cognitive ability, specific cognitive processing abilities, oral language, and achievement. The battery is intended for use from preschool to geriatric ages. The WJ-III is composed of two different test batteries, the WJ-COG and the WJ-ACH and is comprised of 42 subtests.

Tests of Cognitive Abilities. The WJ-COG is comprised of 20 subtests that are designed to measure the seven broad CHC cognitive abilities (i.e., fluid reasoning, crystallized ability, visual processing, auditory processing, short-term memory, long-term retrieval, and processing speed) at the stratum II level as well as several additional clinical and cognitive performance clusters. When combined with the achievement battery, additional CHC academic ability composites are provided. As such, the WJ-COG is the only contemporary test battery designed to measure all hypothesized CHC broad abilities. Each of the WJ-COG broad ability composites is composed of two subtests that measure CHC constructs at the stratum I narrow ability level.

The 14 CHC-based subtests are weighted according to their loadings on the first principal component factor (*g* loading) to form a linear general intellectual ability composite (GIA) based upon the most optimal combination of those weights. Table 3 (p. 185) displays the *g* loadings across the CHC factors on the WJ-COG across the 6 to 18 year old age range. Each of the weights is a combination of the loadings reported in the technical manual (McGrew, Schrank, & Woodcock, 2007) for each of the individual narrow ability subtests that comprise each factor. Across the age ranges, the weight of loadings is relatively stable, with crystallized ability (.23) and fluid reasoning (.19) demonstrating the highest loadings overall. The distribution of weights in Table 3 (p. 185) provides evidence for the invariance of the GIA weighting scheme across the 6 to 18 age range. The fact that the crystallized ability factor has the highest loading on the WJ-COG contrasts with Carroll's (1993) three-stratum model, in which fluid reasoning was hypothesized to have the highest loading on the general factor.

The GIA is purported to reflect *g* at the stratum III level. It should be noted that the previous description refers to the extended battery, a standard battery is also available in which a GIA score can be obtained by administering seven subtests; broad ability composites are not available in the standard battery. As this study is concerned with assessing the validity of the WJ-COG CHC broad factors, all references within this study to WJ-COG variables are made in reference to the extended battery. A graphic representation of the extended structure of the WJ-COG is provided in Figure 2 (p. 216). The WJ-COG is particularly useful as a research tool because the CHC model served as the primary blueprint for its construction. The norming sample (*N* = 8,782) is stratified according to region, community type, sex, and race, and is nationally representative based upon 2000 U.S. census estimates. The WJ-COG has been well reviewed in the technical literature and is commonly utilized in educational and clinical settings to measure a host of cognitive abilities (Sandoval, 2003).

Normative and psychometric data establishing the adequacy of the WJ-COG as a valid and reliable measure of cognitive abilities can be found in the WJ-III technical manual (McGrew & Woodcock, 2001). Mean split-half internal (speeded tests utilized test-retest coefficients) consistency estimates for the WJ-COG variables utilized in this study are provided in Table 4 (p. 186). The average reliability coefficient across all of the indexes is .90. Estimates for all of the individual indexes exceeded .80 with the exception of the visual processing factor (.78). According to Sattler (2008), reliabilities above .80 are preferred for tests used in individual assessment, whereas Salvia and Ysseldyke (2007) adopted a stricter criterion of .90 for utilizing tests for individual decision-making. Test-retest data for the subtests that align with the original WJ cognitive performance structure (e.g., thinking abilities, cognitive efficiency) are provided in the technical manual. Coefficients ranged from .60 (memory for names) to .84 (visual matching), indicating moderate stability across a one year to two year testing interval. However, that information is of limited utility for estimating the temporal stability of the CHC factors on the WJ-COG.

Construct validity for the WJ-COG is established in the technical manual using a series of confirmatory factor analyses (CFA). Additionally, the fit of the WJ-COG three stratum model has been found to be invariant across a number of different samples and clinical conditions (Lock, McGrew, & Ford, 2011). Concurrent validity for the broad factors has been established through a series of joint factor analyses between the WJ-COG and a number of existing cognitive measures (Sanders, McIntosh, Dunham, Rothlisberg, & Finch, 2007). Within these studies, correlations between the GIA and full-scale IQ scores for other intelligence tests have ranged from .67 to .76, providing adequate evidence of convergent validity with external measures purported to measure the same construct.

Tests of Achievement. The WJ-ACH is a comprehensive academic assessment battery designed to measure five academic domains: reading, written language, mathematics, oral language, and academic knowledge. The WJ-ACH is comprised of 22 subtests that combine to provide 17 broad factors and clusters and a total achievement composite score. The WJ-ACH is composed of a standard battery of 11 measures along with an extended battery of seven additional subtests. The WJ-ACH can be interpreted at several levels, which are differentiated according to their generality. At the more discreet level, 11 achievement clusters are composed of one or two corresponding subtests. If the standard battery is administered, than each subtest can be interpreted as a cluster indicator, however if the extended battery is administered the composition of some of the clusters includes an additional subtest. Out of the 11 cluster scores that are possible, five (reading fluency, math calculation skills, math fluency, writing fluency, written expression) remain composed of a single subtest regardless of whether the standard or extended options are utilized. All of the clusters within each of the five achievement domains sampled in the WJ-ACH can be combined to form a corresponding composite broad score (e.g., reading, mathematics, written language, oral language, knowledge). Broad scores from the standard battery each contain two to three subtests whereas, those formed from the extended battery can contain up to four or five indicators. The interpretive structure of the WJ-ACH is outlined in Figure 3 (p. 217). All of the broad academic scores utilized within this study reference the extended battery of the WJ-ACH. The WJ-ACH is widely used to assess students referred for learning disability evaluations due to the fact that cluster scores are provided for all seven of the achievement domains (e.g., basic reading skills, reading comprehension, math calculation skills, math reasoning, written expression, oral expression, listening comprehension) outlined in federal special education learning disability regulations (Mather & Wendling, 2009).

Normative and psychometric data for the WJ-ACH can be found in the WJ-III technical manual (McGrew & Woodcock, 2001). The norming sample (*N* = 8,782) is stratified according to region, community type, sex, and race, and is nationally representative based upon 2001 US census estimates. Split-half internal consistency estimates (test-retest were utilized to assess speeded tests) for the broad achievement factors from the WJ-ACH utilized in this study can be found in Table 4 (p. 186). The mean coefficient across all of the broad achievement indexes and cluster scores is .90. The average coefficients across the 6 to 18 age range for all of the individual measures of focus in this study exceed .80. Test-retest evidence is provided for the WJ-ACH in the technical manual. Correlations were quite high for all of the broad achievement index and cluster scores, with an average coefficient of .96 across a one year assessment interval, with a range of .93 (Written Expression) to .99 (Total Achievement).

Evidence of concurrent validity is provided in the WJ-III technical manual in the form of correlations between various scores on the WJ-ACH and comparable measures from existing achievement batteries. Correlations between the WJ-ACH Total Achievement composite and comparable measures ranged from .65 to .79. Unfortunately, construct validity data are lacking. Although results of a CFA investigation are provided by the test authors to establish the construct validity of the WJ-ACH, that investigation assessed the CHC representation (e.g., quantitative reasoning, long-term retrieval) of the battery and was not a comprehensive investigation of the entire WJ-ACH model reported in the test manual. A subsequent search in the ERIC database using the search terms *Woodcock-Johnson Tests of Achievement*, *Factor Analysis*, and *Validity* yielded no additional independent investigations of the factor structure of the WJ-ACH. It should be noted that the absence of a comprehensive construct validity study for such measures is not uncommon (Kamphaus, 2009). In an independent review by Cizek (2003) the technical properties for the WJ-ACH were lauded, though additional construct validity data was recommended.

Study Variables

A total of 18 cognitive-achievement variables, measured by the WJ-III test battery, were utilized in this study. The variables were all represented in the form of standard score data from representative factor, composite, cluster, and broad scores, on the WJ-III.

Independent variables. Predictor variables (*n* = 8) were selected from the WJ-COG. Independent variable (IV) selection was made on the basis of the most recent iteration of CHC theory (Schneider & McGrew, 2012), resulting in the inclusion of 7 CHC broad stratum II factors and a stratum III general factor represented by the GIA composite. Although several additional stratum II factors (e.g., decision speed, psychomotor speed, domain-specific knowledge) have been hypothesized in the technical literature (e.g., Carroll, 2003), such estimates do not conform to the purported factor structure of the WJ-COG and were not selected for inclusion in the present study.

The predictive model utilized in this study omitted assessment at the stratum I narrow ability level as well as the inclusion of several "clinical" (e.g., thinking ability, cognitive efficiency, working memory) clusters that utilize alternative combinations of CHC-based subtests and non-CHC subtests from the WJ-COG. The inclusion of stratum I measures would introduce serial dependency at multiple levels, inflating the Type II error term (i.e., not rejecting a false null hypothesis). Although it is possible to assess for the effects of stratum I measures in isolation (see Canivez, 2011), primary interpretation at that level is not consistent with any of the PSW models that are based on CHC theory. Furthermore, median internal consistency coefficients meet or exceed a median of .90 for only six of the 20 subtests on the WJ-COG. The inclusion of unreliable measures in a predictive model is a threat to statistical conclusion validity (Kirk, 2013). The three cognitive performance clusters were not included in the predictive model because of questions that have been raised regarding their validity (see Schrank, Miller, Wendling, & Woodcock, 2010), in addition to the fact that the broad abilities that comprise the clusters are accounted for within the parameters of the CHC interpretive model. Definitions from the WJ-COG technical manual for each of the cognitive IVs can be found in Table 5 (p. 187).

Dependent variables. Criterion variables (*n* = 10) were selected from the WJ-ACH. The 2006 final regulations for the Individuals with Disabilities Education Act (IDEA) outline seven achievement domains that may be considered for determining whether a student qualifies for special education and related services under the category of specific learning disability. Those domains are: oral expression, listening comprehension, written expression, basic reading skill, reading comprehension, mathematical calculation, and mathematical reasoning. As evaluative teams are required to consider individual performance in all of those categories, seven of the achievement clusters that correspond directly to the IDEA achievement categories were selected as criterion variables for analysis for the primary research questions associated with analysis of the total sample. It is believed that this selection of achievement variables will provide the most relevant information to practicing school psychologists. For the secondary research questions (questions 3 through 6), the broad scores for reading, mathematics, and written language were utilized as criterion variables. The broad scores were used for questions 3 through 6 because they provide a comprehensive and reliable estimate of several relevant domains of academic achievement while providing for a more parsimonious presentation of results given the number of contrasts that would be required if the seven achievement variables utilized in questions 1 and 2 were retained. Definitions from the WJ-ACH technical manual for each of the cognitive DVs can be found in Table 6. Although the WJ-ACH provides users with several additional fluency, knowledge, and total achievement cluster scores, they were not included in this study as criterion variables because of potential content overlap with the WJ-COG processing speed factor and lack of correspondence with IDEA regulations. Definitions from the WJ-ACH technical manual for each of the achievement DVs can be found in Table 6 (p. 189).

Research Design

This is a descriptive correlational study that utilized a hierarchical or sequential multiple regression (HMR) design for the purposes of examining the incremental validity of the CHC broad factors, represented on the WJ-COG, in predicting achievement outcomes after controlling for the effects of the general factor. The variables were analyzed using the linear regression module in SPSS$^{®}$ version 21. The use of HMR to assess the incremental validity of cognitive variables (e.g., those measured by the WJ-COG) is well established in the technical literature (see Canivez, 2013b).

Explanation vs. prediction. In the social sciences, multiple regression research designs are often described as either being predictive or explanatory in nature. According to Venter and Maxwell (2000), predictive research emphasizes practical applications whereas explanatory research focuses on "achieving a theoretical understanding of the phenomenon of interest," (p. 152). Despite philosophical distinctions, the line between explanation and prediction often is blurred in behavioral research (Hempel, 1965). That is, the results of predictive research have implications for explanation however; they cannot be used alone for such purposes (Schneider, 2008). Because this study is primarily concerned with providing information to practitioners regarding the clinical utility of the WJ-COG, this study is designed to be predictive in nature. Predictive research is critical for establishing the efficacy of interpreting the WJ-COG at the factor level. According to DeGroot, "The criterion par excellence of true knowledge is to be found in the ability to predict the results of a testing procedure. If one knows something to be true, he is in a position to predict; where prediction is impossible, there is no knowledge" (1969, p. 20).

Clinical versus statistical significance. In the current study, a distinction is made between the assessment of statistical significance and clinical or practical significance. Although statistical significance testing is commonly employed in multivariate data analysis, those results provide little information regarding the practical importance of study findings (Thompson, 2007). In contrast to conventional null hypothesis testing, practical significance is typically assessed using effect size estimates and confidence intervals (Stevens, 2009). These estimates place more of an emphasis on the impact of sample size, magnitude of effects, and reliability of the test results. Interpretation in previous incremental validity studies has largely focused on the clinical significance of incremental prediction provided by various factor scores through interpreting the R^2 statistic as an effect size estimate; although no studies could be located which reported corresponding confidence intervals. This study assessed the validity of the WJ-COG from both perspectives. The decision to assess statistical significance in addition to clinical significance was made to avoid creating the perception of selection bias in the presentation of results. As Light, Singer, and Willett (1990) point out, the significance of test results is often found in the eye of the beholder. Therefore, it is believed that presenting all available test results is objectively balanced and is as consistent with best practice guidelines (e.g., Tabachnick & Fiddell, 2007) for reporting HMR research results.

In this study, statistical significance was assessed by interpreting the results of the F test and corresponding incremental F (F_{inc}) statistic. As was previously reviewed, the F test is an analysis of variance test, formatted to multiple regression, which tests the statistical significance of the R^2 statistic while accounting for sample size and the number of predictors in the regression equation. Although it is possible to test for the statistical significance of individual variables via the unstandardized regression coefficients (b) or standardized regression coefficients (beta weights or β), the error term for tests of individual regression coefficients can become inflated in the presence of correlated independent variables. As collinearity often is present when working with variables from intelligence tests, significance tests of the coefficients associated with individual variables will not be conducted and regression coefficients attributable to individual variables will not be reported.

Practical significance was assessed by interpreting the squared multiple correlation (R^2), the incremental squared multiple correlation (ΔR^2) coefficients, and the corresponding confidence intervals, associated with various predictor blocks and individual variables. Because R^2 is a product of the ratio of the regression sum of squares to the total sum of squares it is not impacted directly by collinearity. As an effect size, R^2 indicates the proportion of variance of the dependent variables that is accounted for by the independent variables. As there are no conventional guidelines for interpreting ΔR^2, this study will utilize those recommended by Cohen (1988) for R^2. Thus, ΔR^2 coefficients of .01, .09 and .25 will be interpreted as indicating "small," "moderate," and "large" increases in predictive efficacy associated with the inclusion of additional CHC factors beyond the GIA.

Order of entry. Unlike other regression procedures, variable order in HMR is determined *a priori* according to expected theoretical relationships between the variables and causal priority. According to Cohen et al., (2003), "this is the *only* basis on which variance partitioning can proceed with correlated IVs" (p. 158). The importance of the order in which IVs are entered cannot be overstated as the incremental variance attributed to individual variables or blocks of variables is analyzed after first controlling for the effects of variables that have been previously entered into the regression equation. In general, as more variables are entered into the regression equation (or blocks) there is a diminished return in observed effects on the DV. Because there is consensus within the academic literature that factor/index scores are subordinate to full-scale composites, the GIA will be entered first into the regression equation, followed by various combinations of the CHC indexes or individual CHC factor scores. The order of entry procedures for this study mirror the sequential approach for interpretation that is advocated for most intelligence tests (see Sattler, 2008), including an interpretive handbook written specifically for the WJ-COG (Schrank, Flanagan, Woodcock, & Mascolo, 2002). It should be noted that the authors of the WJ-COG manual (Mather & Woodcock, 2001b) advocated for primary interpretation at the factor level. The purpose of this investigation is to provide information that will help users evaluate the validity of that claim.

Power Analysis

An *a priori* power analysis was conducted to determine the minimum number of cases needed to exhibit adequate power for this study. Analyses were conducted utilizing G*Power 3.1 (Faul, Eerdfelder, Buchner, & Lang, 2009), a software tool for general power analysis. Using a linear fixed model multiple regression design for assessing incremental R^2 increase with seven tested predictors (CHC factors) and eight total predictors (inclusion of the GIA in block one of the regression equation), power equal to .80, and an alpha level of .05, it was determined that a sample size of 153 cases was needed to estimate a medium effect size (f^2 = .09; Cohen, 1988). It is expected that this study will yield small to medium effect sizes based upon the results of previous empirical investigations of the incremental prediction beyond the full-scale score provided by various intelligence test part-scores.

For question 4, adequate sample sizes using established incremental validity assessment procedures were not obtained. Therefore the regression model in such circumstances will be augmented by eliminating the joint entry of all seven CHC IVs in the second block in favor of isolating each variable one at a time. By reducing the degrees of freedom in the resulting design, it is expected that power will be increased. As a result of reducing the degrees of freedom in the design, sample requirements to adequately estimate a medium effect size were reduced to 82 cases. Although this alternative design may eliminate the ability to assess the effect of all CHC factors entered together for some of the research questions, it is believed that their benefit to the overall study is worth such modifications.

Data Analysis Procedures

The specific analyses that were utilized for each of the individual research questions and corresponding hypothesis are discussed in detail below. A summary matrix of the specific hypotheses, predictor, criterion variables, and analytic methods associated with each individual research question is displayed in Figure 4 (p. 218).

Missing data analysis. Missing data are pervasive in behavioral research and particularly a problem when using extant data sets (Allison, 2002). Because this study utilized an archived data file, missing data are anticipated. According to Enders (2010), missing data are defined as the presence of one of more missing variables for an individual case. Although small amounts of missing data in a large data set are not considered to be problematic, large quantities of missing data, even in a large data file, can impact the validity of the test results. The purpose of missing data analysis is to determine whether the missing data is due to an underlying pattern, which may result in a biased parameter estimate (Enders, 2010).

Missing data are usually identified as fitting one of three different patterns that correspond to a classification system first introduced by Rubin (1976). Data are missing at random (MAR) when there is a systematic relationship between one or more measured variables and the probability of missing data. Data that are missing completely at random (MCAR) indicates that observed data points are a random sample of the scores that would have been obtained from a complete sample. Finally, data are missing not at random (MNAR) when the probability of a missing value is associated with the variable that is missing. The most common and preferable method for handling missing data is simply to delete the offending case through either *listwise deletion* or *pairwise deletion* procedures. In listwise deletion, any case that has missing data is deleted prior to analysis, whereas in pairwise deletion cases are eliminated from individual analyses if they are missing only the variable(s) of interest. The default procedure for missing data in SPSS® is listwise deletion, which is the preferred deletion procedure for regression because it allows for a more precise estimate of the error term (Allison, 2002). However, to utilize deletion procedures, missing data must meet the stringent assumptions of MCAR. Deletion with MAR and MNAR data may result in a loss of sample power as well as threatening the validity of correlation matrices (Cohen et al., 2003).

Unfortunately, MCAR is the only pattern that can be assessed systematically in conventional statistical programs, although multiple methods exist for examining missing data. One is to create groups based on missing and non-missing data for a specific variable and then test for mean differences on other variables. However, this procedure can be difficult to utilize with large data sets which have multiple variables. Furthermore, this procedure does not account for potential correlations in the data in which a single mechanism may account for missing data across variables (Enders, 2010). Little's MCAR test (Little, 1988), was designed to accommodate such situations by utilizing an omnibus *t* test approach in which all variables are examined simultaneously. If the test is statistically significant it indicates that the missing data pattern may not be MCAR. This method was utilized in this study to assess the MCAR hypothesis.

Research questions. The specific analyses that will be utilized for each of the research questions and corresponding hypothesis are listed in Figure 4 (p. 218) and described in more detail below.

Question 1: Do the CHC broad ability factors on the WJ-COG provide *statistically* significant incremental prediction of achievement outcomes above and beyond the effects of the general factor? To test if the CHC factors provide for statistically significant incremental prediction of achievement outcomes beyond the effects of the general factor on the WJ-COG, an HMR analysis will be utilized. The main focus of this investigation is to determine whether the contribution of the CHC factors, entered individually and in a joint block, provide for significant incremental achievement prediction beyond chance levels (i.e., $p < .05$). It is necessary to assess both joint and individual factor effects due to the fact that conflicting results are often obtained in regression research by entering the same variables into an equation in different configurations (Pedhazur, 1997). Primary analyses will focus on the results of the *F* test at both stages of model entry. As was previously discussed, individual regression coefficients will not be reported because of the expected threat of collinearity. The inclusion of null hypothesis tests of significance within this study is necessary to provide a balanced assessment of the incremental validity of the WJ-COG. The order of entry for the independent variables will follow typical research on intelligence tests in which the global or IQ score is entered first followed by the lower-order factor scores. Entering the scores in a reverse order would not be consistent with CHC theory and would violate the scientific law of parsimony. The following steps were utilized in this analysis.

1. *Joint block*. An initial regression equation was tested which utilized the GIA in the first block followed by the entry of all seven CHC factors (i.e., fluid reasoning, crystallized ability, visual processing, auditory processing, short-term memory, long-term retrieval, processing speed) simultaneously in a second block. This equation was utilized to assess all seven of the broad achievement (i.e., basic reading skills, reading comprehension, math calculation, math reasoning, written expression, listening comprehension, oral expression) dependent variables in isolation.
2. *Individual factors*. A second regression equation was constructed in which the GIA was entered first followed by a single CHC factor. Each CHC factor was tested against all of achievement variables utilizing the following predetermined order: 1) crystallized ability; 2) fluid reasoning; 3) auditory processing; 4) visual processing; 5) long-term retrieval; 6) short-term memory; 7) processing speed. This order was formulated on the basis of the relative *g* loadings for each of the factors reported in the WJ-COG technical manual (see Table 3).

Question 2: Do the CHC broad ability factors on the WJ-COG provide *clinically* significant incremental prediction of achievement outcomes above and beyond the effects of the general factor? To assess whether the CHC factors provide for clinically significant incremental prediction of achievement after controlling for the effects of the general factor, the results obtained from the procedures described for question 1 were reinterpreted utilizing the R^2 coefficient (with corresponding 95% confidence intervals) as a measure of effect. When the R^2 statistic is utilized as an effect size it can be interpreted as measuring the amount of variation in the DV that is attributable to the IV.

A 95% confidence interval was chosen to be consistent with the adopted criterion in this study for considering test results statistically significant at $p < .05$, and it is often the recommended interval for reporting psychoeducational test results for school psychologists (see Salvia and Ysseldyke, 2007).

Question 3: Is the predictive validity of the CHC factors on the WJ-COG invariant across different levels of schooling? The second part of this study was designed to determine whether the incremental validity of the CHC factors in predicting achievement beyond the general factor on the WJ-COG is impacted by maturational effects. Question 3 is concerned primarily with accounting for the potential effects of maturation and changes in cognitive-achievement relations that may be the result of curricular or developmental changes across the lifespan. To assess for maturational effects, the total sample was divided by grade into several groups associated with various levels of schooling. The sample was divided into an “elementary” group (grades K through 5); “intermediate” group (grades 6 through 8); and a “secondary” group (grades 9 through 12). The following steps were utilized to assess the incremental validity of the grade groupings.

1. *HMR analysis*. The incremental validity of each group was assessed utilizing the multi-step HMR procedures described for questions 1 and 2, whereby the CHC factors were assessed as a joint block (with the broad achievement scores for reading, math, and writing serving as criterion variables). Because the remaining parts of this study were concerned primarily with conditions that may potentially mediate incremental validity estimates, interpretation will focus

primarily on the proportions of variance explained via the R^2 statistic and the corresponding confidence intervals.

2. *Testing for homogeneity across groups*. To estimate whether the incremental validity of the CHC factors is invariant across various levels of schooling, the obtained R^2 estimates for the three groups were converted to proportions and analyzed using a chi-square goodness of fit test utilizing the R^2 estimates obtained from the total sample as reference parameters. Total variance for the model was summed to 100 and divided into three categories: achievement variance attributable to the GIA, variance attributable to the CHC factors, and unpredictable variance (which encompasses an unspecified combination of error variance and factors not specified within the prediction model). If a chi-square test is significant, it indicates that the obtained estimates for a particular group differ from that which would be expected from the total standardization sample, indicating that school level is a moderating variable. Non-parametric statistics were utilized to assess for group-level effects due to the fact that the proportions being utilized for the analyses do not meet the assumptions for conventional parametric tests (e.g., ANOVA, *t* test).

Question 4: Is the predictive validity of the general factor attenuated by significant levels of inter-factor variability on the WJ-COG? To assess whether inter-factor variability among the CHC broad factor scores impacts the predictive validity of the general factor on the WJ-COG, the total sample was divided into three groups based upon level of observed differences between fluid reasoning (Gf) and crystallized ability (Gc). The decision to utilize Gf-Gc variability as the criterion marker for this research question was based upon three factors. First, Gf and Gc have the highest *g* loadings on the WJ-COG, accounting for over 40% of the total loadings on the GIA. Thus, it is believed that significant differences observed on these factors will have the greatest chance of destabilizing the resulting general factor estimate. Secondly, the validity evidence compiled by Carroll (1993) for the hypothesized stratum II broad abilities was strongest for Gf and Gc. Significant questions were raised about the nature and composition of all of the remaining factors. Finally, the Gf-Gc dichotomy often serves as a proxy for the long standing verbal-nonverbal debate in the technical literature. Significant differences between verbal and non-verbal performance on intelligence tests have long been associated with various forms of pathology (see Kaufman & Lichtenberger, 2006 for a comprehensive review). In response, researchers (e.g., Hale & Fiorello, 2004) often encourage practitioners to abstain from interpreting the general factor when significant discrepancies between verbal and non-verbal performance are observed.

Three discrepancy groups were created and each case was coded with a binary variable indicating whether the specified level of Gf-Gc discrepancy was present. 15 point, 23 point, and 30 point groups were categorized. The choice of discrepancy levels was not made arbitrarily as the levels correspond to various levels of deviation from the mean in the theoretical normal distribution. Unfortunately, the preferred method of assessing significance levels which correspond to base rates observed in the standardization sample (e.g., 5%, 10%, 15%) is not possible because that information is not reported in the WJ-COG technical manual or corresponding computer scoring system. However, utilizing the distribution of the total subsample of cases in which both a Gf and a Gc score were obtained (n = 2,817), the discrepancy intervals fall at roughly the 25th, 10th, and 5th percentiles. These values are commensurate with levels that have been assessed in other studies that examined the effect of inter-factor variability on other measures (e.g., Freberg, Vandiver, Watkins, & Canivez, 2008), and are consistent with levels of statistical significance commonly reported in test manuals.

The same procedures described for question 3 were utilized to assess question 4 with one notable departure. Because the normal distribution results in only 3.4% of observations within the sample with a 30 point or more Gf-Gc discrepancy, the corresponding discrepancy group will not have adequate power to estimate a moderate to large effect size (e.g., .09 to .20) if the joint CHC factor regression equation is utilized. To increase power, the HMR model for the 30 point discrepancy group was limited to assessing the effects of each CHC factor individually. The proportions accounted for by each CHC factor were then summed to provide an estimate of overall effects attributable to the stratum II factors. In the secondary chi-square analysis, the cases that do not present with a significant discrepancy (i.e., Gf-Gc differences of less than 15 points) served as a control group and were utilized as a reference marker for the expected variance distributions at each level of analysis.

Question 5: Does Spearman's law of diminishing returns (SLODR) impact the predictive validity of the general factor on the WJ-III COG? To assess the potential effects of SLODR on the predictive validity of the WJ-COG, incremental validity across groups was assessed using the same procedures for questions 3 and 4. The groups were differentiated according to overall level of cognitive ability as estimated by the GIA score. Specifically, a "below average" group composed of cases was created with GIA scores of 80 and below, "average ability" group (81 to 119), and "above average" group

(120 and above). Although there is the potential for the below average and above average groups to be under-sampled, it was anticipated that both groups would contain enough cases to adequately estimate a moderate to large effect size using the joint factor regression equation. To test the SLODR hypothesis, the variance coefficients obtained from HMR analysis of the average subsample served as parameter estimates for the below average and above average groups. A chi-square statistic exceeding the established critical values for both the above average and below average groups would be expected for the effects of SLODR to be demonstrated.

Question 6: Does the use of a differential weighting scheme enhance the validity of the WJ-COG factor structure in predicting norm-referenced reading, math, and writing outcomes when compared to estimates that have been obtained from other intelligence tests using similar methods of variance partitioning on commercial standardization samples? Because the R^2 coefficient can be interpreted as an effect size estimate and is the most commonly reported statistic in incremental validity studies of intelligence tests, the variance coefficients obtained in this study can be synthesized with the results obtained from studies using similar research designs. Because all of the other intelligence tests that have been subjected to incremental validity investigations using HMR utilize the same weighting procedure (i.e., linear average) to develop their representative full-scale composites, the results obtained from this investigation can be compared to results from other intelligence tests using fixed effects meta-analytic techniques. By utilizing fixed effects meta-analytic procedures, evidence can be compiled to determine whether the weighting procedure and/or the purported factor structure of the WJ-COG is able to account for greater levels of reading, math, and writing achievement when compared to other intelligence tests. To be included within the synthesis studies must meet the following criteria.

1. The design must be an incremental validity investigation utilizing HMR.

2. The study must utilize a regression equation with the general factor score entered first followed by relevant factor scores jointly entered into the second block.

3. To maintain consistency, only the criterion variables, of reading, mathematics, and writing composites will be obtained.

4. The study must utilize a linked norm-referenced achievement test subsample from a commercial standardization project for analysis.

Unfortunately, the primary analysis (i.e., does the WJ-COG model account for more achievement variance) was predominately descriptive in nature.

The effect size estimates for each of the included studies were inversely weighted according to the procedures for correlational effect sizes outlined in Hedges and Olkin (1985) so that effect size estimates could be synthesized into grand values (i.e., $\overline{R^2}$, $\overline{\Delta R^2}$). To utilize these procedures, each R^2 and ΔR^2 coefficient was transformed into a z-score prior to being weighted (i.e., multiplied) by a variance term ($W_k = N - 3$) that took into account the sample size for each study. The $\overline{R^2}$ and $\overline{\Delta R^2}$ values obtained in the meta-analysis of incremental validity studies of other intelligence tests were utilized as parameter estimates for comparison with the WJ-COG total sample estimates from the current study. The statistical significance of differences between proportions was assessed using the chi-square goodness of fit test. Tests of homogeneity/heterogeneity (e.g., Q and I^2) were conducted to determine if a fixed effects model could be assumed. In a fixed effects model, differences between effect sizes from one study to another are attributed to variation that would be expected from sampling error.

The *Q* test follows a chi-square distribution and is computed by summing the squared deviations of each study effect size estimate from the grand effect size estimate. If a *Q* test is significant, it indicates that a random effects model, in which effect size differences between studies are thought to be the result of within-study variability is assumed. Conversely, a non-significant *Q* test indicates it is appropriate to synthesize effect sizes for the studies, despite between-study differences that may be observed (e.g., composite weighting schemes, factor structure, theoretical orientation).

The I^2 statistic describes the percentage of total variation across studies that is due to heterogeneity rather than chance and is often presented as a supplemental descriptive statistic for *Q*. Higgins and Thompson (2002) suggested the following guidelines for interpretation: I^2 = 25% (small heterogeneity), I^2 = 50% (medium heterogeneity), and I^2 = 75% (large heterogeneity). Negative I^2 values are automatically set to zero.

The use of a meta-analytic format has the potential to contextualize the results obtained in this study within the broader incremental validity literature and may provide practical information to users regarding the efficacy of various scores from contemporary intelligence tests in predicting outcomes on norm-referenced achievement tests.

Permissions and Institutional Approval

Permission to utilize the WJ-III standardization data for this study was obtained from the Woodcock-Munõz Foundation (WMF). A copy of the WMF limited use agreement is provided in Appendix A (p. 222). Institutional approval was obtained from the Chapman University Institutional Review Board (IRB) on 6/27/13. A copy of the IRB approval notice for this study is provided in Appendix B (p. 224).

Summary

The current study was designed to assess the predictive validity of the CHC factors on the WJ-COG in accounting for norm referenced achievement outcomes after controlling for the effects of the general factor. The incremental validity of the WJ-COG was appraised across six research questions. A hierarchical regression design was utilized to assess standard score data obtained from individuals who participated in the original WJ-III standardization project. Data analysis was focused primarily on interpreting regression coefficients as effect sizes using conventional guidelines (e.g., Cohen, 1988). Secondary analysis consisted of evaluating the statistical significance of prediction model fit using the chi-square statistic. This study was designed to be predictive in nature in order to provide useful information to practitioner's who utilize the WJ-COG in clinical practice.

Chapter IV: Results

Univariate Descriptive Statistics and Evaluation of Assumptions

The standard scores of the Woodcock-Johnson III (WJ-III), which have a mean of 100 and a standard deviation of 15, were used in this study. The means, standard deviations, skewness, and kurtosis statistics for all of the WJ-III cognitive and achievement variables are listed in Table 7 (p. 191). The mean (99.99 to 101.38) and standard deviation ranges (14.62 to 16.08) for the cognitive and achievement variables generally reflect values that would be expected for normally distributed standard score variables. Skewness values for all the variables were between -.41 to .11. According to Bulmer (1979), skewness statistics between -.50 and .50 approximate normally distributed symmetry. However, there are several statistical methods for assessing the relative normality of a sample distribution in addition to the examination of skewness and kurtosis statistics (Stevens, 2009). The results of those assessments are described below.

In SPSS® it is possible to test the normality of a sample distribution using the Kolmogorov-Smirnov and Shapiro-Wilk tests, which empirically assess frequency and probability plot distributions. Tabachnick and Fiddell (2007) recommended the use of conservative alpha levels (e.g., $p < .001$) for interpreting the significance of individual tests of normality. The results of the Kolmogorov-Smirnov and Shapiro-Wilk tests are listed in Table 8 (p. 192). Kolmogorov-Smirnov values ranged from .02 to .04 across the variables, with six coefficients meeting or exceeding the established alpha level for statistical significance. Shapiro-Wilk values ranged from .98 to .99, 14 of which were considered to be statistically significant. Tabachnick and Fiddell (2007) warned that it is common to obtain significant coefficients on tests of normality with samples that include more than 200 cases. As a solution, they recommended visually examining histogram distributions and probability plots for individual variables when conducting multivariate analysis. Visual inspection of these indicators was not significant; indicating that the standard scores obtained for all of the WJ-III variables generally reflected a normative distribution. It should be noted that un-grouped normally distributed variables also meet assumptions for linearity and homoscedasticity (variance is evenly distributed throughout variables). Additionally, inspection of the residual plots of the data indicated that the regression models utilized in this study met the assumptions for homoscedasticity of residuals.

Outlier assessment. The WJ-III scoring program reports standard scores for individual variables that range from zero to 200. Minimum and maximum values were examined for each variable to detect potential calculation errors. Obtained scores for all of the variables fell within the specified WJ-III standard score range. According to Stevens (2009), any cases that fall two or more standard deviations from the mean indicate the presence of potential outliers in the dataset. In multiple regression, outliers are problematic because they exert leverage on the underlying regression coefficients. The study variables had between 99 and 210 (*M* = 169) cases that fell outside of the two standard deviation band. Statistical measures used to identify multivariate outliers are leverage, discrepancy, and influence. Leverage refers to distance of a case from the centroid of the variable means and is typically tested through interpretation of hat values (measure of distance from the centroid point of means) via the Mahalanobis distance test. Discrepancy refers to the degree to which a case is in line with the others and is assessed by examining residual plots for outliers. Influence is the product of leverage and discrepancy and is assessed by examining the impact on regression coefficients, once outlier cases are removed from the analysis via Cook’s distance tests. In test models that utilized broad achievement scores as criterion variables, leverage and discrepancy tests were significant, whereas Cook’s distance values fell below thresholds that have been recommended as indicating significance (i.e., less than one; Tabachnick & Fiddell, 2007). According to Fox (1991), the decision as to whether to delete outlier variables from a regression analysis should be based primarily upon the presence of significant Cook’s distance values and not the results of leverage or discrepancy tests interpreted in isolation. On the basis of these evaluations, it was determined that no cases met the criteria for exclusion from the current analyses.

Collinearity diagnostics. As was previously discussed in Chapter II, collinearity occurs when correlated predictor variables are entered simultaneously into a regression equation. Collinearity may result in unstable regression coefficients due to inflated standard errors. Bivariate correlation coefficients between the independent variables (IVs) are provided in Table 9 (p. 193). The coefficients are moderate to strong, ranging from .25 to .84. All of the coefficients are statistically significant ($p < .01$, two-tailed), indicating that collinearity was likely. Diagnostics were run on test models utilizing the broad achievement scores as criterion variables in the regression equation. To review, collinearity is typically assessed by analyzing one or more of several indices (e.g., Tolerance, Variance Inflation Factor, Condition values). Tolerance refers to the variance in an IV that is independent of other IVs. Although the default Tolerance cutoff value in SPSS® is .0001, values as high as .10 indicate that there may be serious problems with serial dependency (Cohen, Cohen, West, & Aiken, 2003). Variance Inflation Factor (VIF) is the inverse of Tolerance and indicates the amount of error for each regression coefficient that is increased relative to the null hypothesis that all IVs are uncorrelated. VIF values greater than or equal to 10 commonly are interpreted as indicating significant collinearity. Finally, the Condition Number, also known as kappa (κ) is the square root of the ratio of the largest eigenvalue to the smallest eigenvalue, which is obtained from principal components analysis of the IVs. Pedhazur (1997) recommended examining condition numbers and corresponding variance decomposition proportions simultaneously. For diagnosing collinearity, Belsey (1991) suggested that large condition values (e.g., ≥ 30) that also have variance proportions of .50 or above, for two or more variables, be interpreted as problematic. Tolerance, VIF, and Condition values are provided for each predictor variable in Table 10 (p. 194). Across the reading, math, and writing models, all indicators were significant for at least two variables, indicating that collinearity was an area of concern. Although not a threat to the R^2 statistic, the results of the collinearity tests validate the decision not to interpret or report the regression coefficients obtained in this study.

Missing Data Analysis

Missing data analysis was completed to determine the amount of missing data in the sample utilized in the current analyses. Summary missing data statistics for each of the 18 WJ-III variables are listed in Table 11 (p. 195). Across the variables utilized, 15% to 55% of the cases contained missing data, whereas 36% (n = 1,702) of the cases had complete data on all of the variables. Little's test for Missing Completely at Random (MCAR) was statistically significant across the sample χ^2 (2,581) = 3593.42, $p < .001$, indicating that the MCAR hypothesis may not be tenable. Tests have not been developed to directly assess whether data are missing at random (Enders, 2010). However, because the data were gathered from a variety of test sites across the United States, and collected at different times, there is no reason to believe that there is any particular systematic mechanism from the standardization study design that is related to the missing data.

Of the methods that are available for treating missing data, only maximum likelihood estimation and multiple imputation have been well received in the technical literature (Schafer & Graham, 2002). Unlike maximum likelihood estimation, multiple imputation (MI) procedures allow a researcher to estimate and replace missing values. The MI method utilizes maximum expectation procedures to generate multiple datasets. The corresponding parameter estimates are then pooled to generate replacement values. However, it is important to remember that the primary goal of statistical analysis is to estimate population parameters. Enders (2010) recommended comparing obtained parameter estimates with those that are generated from a validated estimation technique (e.g., multiple imputation) prior to utilizing imputed values in data analysis. If negligible differences between the two sets of estimates are observed, it would make little sense to analyze fabricated values as the current study was designed to be predictive in nature with observed-level data.

Parameter estimates obtained via MI were compared to those calculated from the existing sample using the Hedge's *g* statistic. Hedge's *g* is an effect size estimate, which differs from a conventional means difference ratio in that it utilizes a pooled standard deviation for the denominator. A pooled standard deviation is preferred because it accounts for differences in size and variation across samples and is a better estimate of variability within the population (Grisson & Kim, 2005). Post-imputation descriptive statistics for the study variables are provided in Table 12 (p. 196). Hedge's *g* values ranged from -.030 to .020, indicating that MI parameter estimates were commensurate with those obtained from the original sample.

Because the MI model failed to provide a more precise estimate of the population, and preliminary analysis indicated that the current sample yielded adequate power, it was determined that the use of imputed values would only serve to obfuscate any inferences that could be made from the current study that potentially were relevant to clinical practice. It is also worth noting that the Tests of Cognitive Abilities (WJ-COG) was designed to support the use of selective testing by users (K. S. McGrew, personal communication, November 22, 2010). Thus, the missing data patterns observed in the present sample reflect common assessment practices of school psychologists who utilize the WJ-COG in clinical practice settings. Furthermore, listwise deletion has fared well in generating unbiased estimates of sample estimates of R^2 in simulation studies (see Brockmeier, Kromrey, & Hines, 1998). On this basis, data analysis was completed utilizing listwise deletion procedures (i.e., cases with missing data for any of the variables of interest were eliminated from regression models).

Incremental Validity of CHC Broad Factors in Predicting IDEA Outcomes

The first part of this study was designed to assess the overall incremental validity of the Cattell-Horn-Carroll (CHC) factors on the WJ-COG when utilized to predict norm-referenced Tests of Achievement (WJ-ACH) variables aligned with specific learning disability criteria found in the Individuals with Disabilities in Education Act (IDEA). Hierarchical multiple regression analyses (HMR) were conducted for each of the seven IDEA achievement categories, utilizing a predictive model with the General Intellectual

Ability (GIA) composite entered first in the regression equation followed by joint and individual combinations of each of the seven CHC broad factor scores measured by the WJ-COG. Tables 13 through 19 (pp. 197-203) report the *F* ratio, F_{inc}, R^2, and ΔR^2 statistics, as well as the corresponding R^2 confidence intervals, for the regression models utilized to predict each of the WJ-ACH variables. $R^2/\Delta R^2$ values reflect the proportion of variance accounted for as each successive predictor variable was entered into the regression for each dependent variable. The *F* ratio assesses the statistical significance of the obtained R^2 estimates using an analysis of variance (ANOVA) model. A significant *F* ratio indicates that a prediction model accounts for significant criterion variable variance beyond chance levels. It is important to note that in an HMR analysis, *F* ratios for secondary blocks reflect the significance of the full prediction model (i.e., first block and secondary block variables simultaneously) whereas, the F_{inc} statistic evaluates the significance of secondary variables, while controlling for the effects of variables already entered into the regression equation.

Basic reading skills. The results from the HMR model utilized to predict basic reading skills (n = 2,129) are reported in Table 13 (p. 197). Both the GIA, F (7, 2120) = 1802.68, $p < .05$, and the CHC factors, F_{inc} (7, 2120) = 15.51, $p < .05$, entered together in a joint block, accounted for statistically significant increments of reading skill variance. Additionally, incremental *F* ratios for the crystallized ability, fluid reasoning, auditory processing, visual processing, and short-term memory factors all were statistically significant. At a practical level, the GIA accounted for 46% of basic reading skill variance, as indicated by the R^2 value of .46, 95% CI [.43, .49], representing large effects. As a joint block, the CHC factors accounted for a small portion of incremental variance, after controlling for the effects of the GIA as evidenced by the ΔR^2 value of .03. Variance increments for individual CHC factors ranged from .01 (crystallized ability) to .02 (fluid reasoning), with the remaining broad abilities accounting for no additional reading variance. Overall, the GIA accounted for 94% of the predictable variance from the regression model.

Reading comprehension. The results from the HMR model utilized to predict reading comprehension (n = 1,813) are reported in Table 14 (p. 198). Both the GIA, F (7, 1806) = 2155.69, $p < .05$, and the CHC factors, F_{inc} (7, 1806) = 38.87, $p < .05$, entered together in a joint block accounted for statistically significant increments of reading comprehension variance. Additionally, incremental F ratios for all of the CHC factors, with the exception of auditory processing, were statistically significant. At a practical level, the GIA accounted for 54% of reading comprehension variance overall ($R^2 = .54$, 95% CI [.51, .57]), representing large effects. As a joint block, the CHC factors accounted for a small portion of incremental variance, after controlling for the effects of the GIA, as evidenced by the ΔR^2 value of .06. Variance increments for individual CHC factors ranged from .01 (fluid reasoning, short-term retrieval) to .05 (crystallized ability), with auditory processing, visual processing, long-term retrieval, and processing speed accounting for no additional reading variance. Overall, the GIA accounted for 90% of the predictable variance from the regression model.

Math calculation skills. The results of the HMR model utilized to predict math calculation skills (n = 2,106) are reported in Table 15 (p. 199). Both the GIA, F (7, 2099) = 842.03, $p < .05$, and the CHC factors, F_{inc} (7, 2099) = 27.40, $p < .05$, entered together in a joint block accounted for statistically significant increments of calculation variance.

Additionally, incremental *F* ratios for crystallized ability, auditory processing, and processing speed, were statistically significant. At a practical level, the GIA accounted for 29% of math calculation variance overall as reflected in the R^2 value of .29, 95% CI [.25, .32], indicating large effects. As a joint block, the CHC factors accounted for a small portion of incremental variance, after controlling for the effects of the GIA as evidenced by the ΔR^2 value of .06. Individually, only the processing speed factor accounted for additional variance ($\Delta R^2 = .04$) when entered alone into the regression equation. It should be noted that differences observed between total variance estimates for joint factor blocks and the summed total of variance accounted for by individual CHC factors is a result of where variables are entered into regression equation as well as rounding errors. Overall, the general factor accounted for 83% of the predictable variance from the regression model.

Math reasoning. The results of the HMR model utilized to predict math reasoning ($n = 2{,}127$) are reported in Table 16 (p. 200). Both the GIA, $F(7, 2120) = 2482.89$, $p < .05$, and the CHC factors, $F_{\text{inc}}(7, 2120) = 15.74$, $p < .05$, entered together in a joint block accounted for statistically significant increments of reading comprehension variance. Additionally, incremental *F* ratios for all of the CHC factors, with the exception of visual processing, long-term retrieval, and processing speed, were statistically significant. At a practical level, the GIA accounted for 54% of math reasoning variance overall as reflected by the R^2 value of .54, 95% CI [.51, .57], representing large effects. As a joint block, the CHC factors accounted for a small portion of incremental variance, as evidenced by the ΔR^2 value of .02. The crystallized ability, fluid reasoning, and auditory processing factors each accounted for an additional 1% of predictive variance. The incremental prediction of the remaining CHC factors was negligible. Overall, the GIA accounted for 96% of the predictable variance from the regression model.

Written expression. The results of the HMR model utilized to predict written expression (n = 2,063) are reported in Table 17 (p. 201). Both the GIA, F (7, 2056) = 1430.80, $p < .05$, and the CHC factors, F_{inc} (7, 2056) = 31.16, $p < .05$, entered together in a joint block accounted for statistically significant increments of writing variance. Additionally, incremental F ratios for the crystallized ability, fluid reasoning, auditory processing, and processing speed factors, were statistically significant. At a practical level, the GIA accounted for 41% of writing variance overall as reflected by the R^2 value of .41, 95% CI [.38, .44], representing large effects. As a joint block, the CHC factors accounted for a small portion of incremental variance, after controlling for the effects of the GIA, as evidenced by the ΔR^2 value of .06. Variance increments for individual CHC factors ranged from .01 (fluid reasoning) to .02 (processing speed), with the remaining factors accounting for no additional incremental prediction. Overall, the GIA accounted for 87% of the predictable variance from the regression model.

Oral expression. The results of the HMR model utilized to predict oral expression (n = 2,126) are reported in Table 18 (p. 202). Both the GIA, F (7, 2119) = 1906.55, $p < .05$, and the CHC factors, F_{inc} (7, 2119) = 242.35, $p < .05$, entered together in a joint block accounted for statistically significant increments of oral expression variance. Additionally, incremental F ratios for crystallized ability, fluid reasoning, visual processing, short-term memory, and processing speed, were significant. At a practical level, the GIA accounted for 47% of oral expression skill variance (R^2 value of .47, 95%

CI [.44, .50]), representing large effects. As a joint block, the CHC factors accounted for a large portion of incremental variance, after controlling for the effects of the GIA as evidenced by the ΔR^2 value of .23 (23%). Incremental R^2 values of this magnitude have rarely been reported in prior incremental validity research. Interestingly, almost all of the additional prediction at the factor level was attributable to the crystallized ability factor ($\Delta R^2 = .23$). Variance increments for other CHC factors ranged from .01 (visual processing) to .04 (short-term retrieval. Whereas the GIA accounted for large effects in predicting oral expression, the combination of CHC factor scores and/or the inclusion of the crystallized ability factor alone accounted for almost a third of the variance from the entire regression model.

Listening comprehension. The results of the HMR model utilized to predict listening comprehension ($n = 2{,}130$) are reported in Table 19 (p. 203). Both the GIA, F (7, 2123) = 2753.28, $p < .05$, and the CHC factors, F_{inc} (7, 2123) = 37.91, $p < .05$, entered together in a joint block accounted for statistically significant increments of listening comprehension variance. Additionally, incremental F ratios for crystallized ability, fluid reasoning, visual processing, and short-term retrieval, were statistically significant. At a practical level, the GIA accounted for 56% of listening comprehension variance overall as reflected by the R^2 value of .56, 95% CI [.54, .59], representing large effects. As a joint block, the CHC factors accounted for a small portion of incremental variance, after controlling for the effects of the GIA, as evidenced by the ΔR^2 value of .05. Variance increments for individual CHC factors ranged from .01 (fluid reasoning, short-term retrieval) to .02 (crystallized ability), with the remaining factors accounting for no additional incremental prediction. Overall, the general factor accounted for 92% of the predictable variance accounted for by the regression model.

Summary. Across the seven HMR models utilized to predict basic reading, reading comprehension, math calculation, math reasoning, written expression, oral expression, and listening comprehension skills on the WJ-ACH, the general factor (as estimated through the GIA) accounted for 29% to 56% (*Mdn* = 47%) of the dependent variable variance, and 66% to 96% of the predictable achievement variance in the models. The R^2 values that corresponded to those variance increments all indicate large effects using Cohen's (1988) interpretive guidelines. CHC broad factors entered jointly into the second block of the regression equations accounted for 3% to 23% (*Mdn* = 6%) of incremental variance, and 6% to 61% of the predictable achievement variance in the models. The ΔR^2 values that corresponded to those variance increments indicated small to large effects using Cohen's guidelines for interpretation. The incremental variance coefficients attributed to individual CHC factors ranged from 0% to 23%, with only the crystallized ability factor in the oral expression model accounting for more than 5% of achievement variance. Although significance tests (e.g., ANOVA) suggest that the CHC factors on the WJ-COG contribute incremental validity beyond the effects of the general factor, effect size estimates were not significant at the practical level of interpretation for all of the regression models, with an exception in the domain of oral expression.

Effects of School Level on the Predictive Validity of the WJ-COG Model

In the second part of this study, the total sample was divided into three grade-level subgroups to determine if level of schooling resulted in predictive validity coefficients that differed from those obtained from analyses of the total study sample.

Variance coefficients as well as incremental variance proportions for the HMR analyses across multiple levels of schooling are provided in Table 20 (p. 204).

Broad reading. In predicting broad reading, the GIA accounted for 48% to 64% of the dependent variable variance, and 89% to 91% of the predictable achievement variance in the models; the *R²* values that corresponded with those variance increments all represented large effects. CHC broad factors entered jointly into the second block of the broad reading regression equations accounted for 5% to 7% of additional variance in the models. The Δ*R²* values that corresponded to those variance increments all represented small effects. The results of the secondary chi-square analyses comparing the obtained variance estimates to those that were calculated from the total sample are provided in Table 21 (p. 206). Goodness of fit tests for differences observed between the variance coefficients obtained from the HMR analyses of the primary, χ^2 (2) = 2.69, *p* = .26, intermediate, χ^2 (2) = .20, *p* = .90, and secondary subsamples, χ^2 (2) = 3.55, *p* = .17, and parameter estimates from the total sample were not statistically significant. Despite the chi-square test results, the predictive effects of the general factor increased systematically as level of schooling increased. At the secondary level, the GIA accounted for almost two-thirds of norm-referenced reading performance as reflected by the *R²* value of .64, 95% CI [.60, .69], the largest effect size estimate obtained in the current study.

Broad mathematics. In predicting broad mathematics, the GIA accounted for 45% to 49% of the dependent variable variance, and 92% to 94% of the predictable achievement variance in the models; the *R²* values that corresponded with those variance increments all represented large effects. CHC broad factors entered jointly into the second block of the broad mathematics regression equations accounted for 3% to 4% of additional variance in the models. The ΔR^2 values that corresponded to those variance increments all represented small effects. Goodness of fit tests for differences observed between the variance coefficients obtained from the HMR analyses of the primary, χ^2 (2) = .04, p = .98, intermediate, χ^2 (2) = .37, p = .83, and secondary subsamples, χ^2 (2) = .43, p = .81, and parameter estimates from the total sample were not statistically significant.

Broad written language. In predicting broad written language, the GIA accounted for 41% to 55% of the dependent variable variance, and 89% to 95% of the predictable achievement variance in the models. The R^2 values that corresponded to those variance increments all represented large effects. CHC broad factors entered jointly into the second block of the broad mathematics regression equations accounted for 3% to 6% of additional variance in the models. The ΔR^2 values that corresponded to those variance increments all represented small effects. Goodness of fit tests for differences observed between the variance coefficients obtained from the HMR analyses of the primary, χ^2 (2) = 1.05, p = .59, intermediate, χ^2 (2) = .22, p = .90, and secondary subsamples, χ^2 (2) = .3.56, p = .17, and parameter estimates from the total sample were not statistically significant. Despite the chi-square test results, the predictive effects of the general factor increased systematically as level of schooling increased. At the secondary level, the GIA accounted for over half of norm-referenced writing performance as reflected by the R^2 value of .55, 95% CI [.50, .60].

Summary. Across reading, mathematics, and writing outcomes, positive developmental trend was evidenced for the general factor, as can be seen through an inspection of the variance coefficients at each schooling level. In reading, the variance accounted for by the general factor increased linearly from the primary group to the secondary group by approximately 16%. In writing, the predictive validity of the general factor increased in similar fashion by approximately 14%. Growth in mathematics was less consistent. Although there was a 4% increase from the primary group to the intermediate group, the general factor accounted for 2% less math variance from the intermediate to the secondary group. When entered jointly in the second block of the regression equation, the CHC factors provided 3% to 7% of additional predictive variance across the WJ-ACH outcome indicators. Overall, these results do not indicate level of schooling is a moderating variable for predicting achievement outcomes with cognitive variables on the WJ-COG.

Effect of Gf-Gc Variability on the Incremental Validity of the WJ-COG

To assess whether inter-factor variability impacted the incremental validity of the broad factors on the WJ-COG, the total study was divided into three subgroups based upon the observed level of variability between crystallized ability and fluid reasoning. The effects of a 15 point difference, 23 point difference, and 30 point difference, as well as a non-significant difference control group, were investigated separately for the broad reading, mathematics, and written language scores on the WJ-ACH. The results of those HMR analyses are displayed in Tables 22 through 24 (pp. 207-209). Secondary chi-square goodness of fit tests was also conducted, which assessed the significance of differences between variance proportions obtained from the control group and those found in each of the discrepancy groups. The results of those tests, across reading, math, and writing outcomes, are displayed in Table 25 (p. 210).

Broad reading. Variance coefficients for the HMR analyses testing the effects of successive levels of Gf-Gc discrepancies on the predication of broad reading are presented in Table 22 (p. 207). In the analysis of the non-significant control group (*n* = 1,529), the GIA accounted for 55% of broad reading variance whereas the CHC factors, entered jointly, provided 4% of additional predictive power. The predictive power of the GIA remained somewhat consistent across all of the discrepancy groups, with R^2 values ranging from .53 to .60, all representing large effects. However, the variance accounted for by the CHC factors increased linearly as the level of inter-factor variability increased. At the 15 and 23 point difference levels, the factors provided 11% of additional variance, which represented moderate effect size increases. In the HMR analysis of the 30 point group (*n* = 65), the variance coefficients for each of the CHC factors, when they were entered into the second block of the regression equation, ranged from 0% to 14%. The incremental coefficients for the crystallized ability ($\Delta R^2 = .12$) and fluid reasoning ($\Delta R^2 =$ .14) factors reached moderate effect size levels. When summed to estimate the effects of entering all of the variables together jointly, the CHC factors accounted for an additional 33% of broad reading variance, which is a large estimated effect. The observed differences between the discrepancy groups and the control group were statistically significant at each level of Gf-Gc variability.

Broad mathematics. Variance coefficients for the HMR analyses testing the effects of successive levels of Gf-Gc discrepancies on the predication of broad mathematics are presented in Table 23 (p. 208). In the analysis of the non-significant control group (*n* = 1,561), the GIA accounted for 46% of broad reading variance whereas the CHC factors, entered jointly, provided 3% of additional predictive power. The predictive power of the GIA remained fairly stable as variability increased; the obtained R^2 coefficients (.39 to .48) all represented large effects. The variance accounted for by the CHC factors increased linearly as the level of inter-factor variability increased, with ΔR^2 values ranging from .03 to .09, representing small to moderate incremental effects. The goodness of fit test for the 30 point group was significant, χ^2 (2) = 15.17, *p* = .001, whereas tests on the 15 and 23 point groups failed to exceed critical levels.

Broad written language. Variance coefficients for the HMR analyses testing the effects of successive levels of Gf-Gc discrepancies on the predication of broad written language are presented in Table 24 (p. 209). In the analysis of the non-significant control group (*n* = 1,528), the GIA accounted for 46% of written language variance whereas the CHC factors, entered jointly, provided 4% of additional predictive power. The predictive power of the GIA remained consistent across all of the discrepancy groups with R^2 values ranging from .44 to .49, all representing large effects. However, the variance accounted for by the CHC factors increased linearly as the level of inter-factor variability increased. At the 15 and 23 point difference levels, the factors provided 8% of additional variance, which represented small effect size increases. In the HMR analysis of the 30 point group (*n* = 68), the variance coefficients for each of the CHC factors when they were entered into the second block of the regression equation ranged from 0% to 5%. However, when summed to estimate the effects of the entering all of the variables together jointly, the CHC factors accounted for an additional 11% of writing variance, a moderate effect size increase. Additionally, the goodness of fit test for the 30 point group was statistically significant, χ^2 (2) = 12.84, *p* = .002.

Summary. Across increasing levels of Gf-Gc variability, the GIA accounted for 39% to 60% (*Mdn* = 47%) of achievement variance on the WJ-ACH. All of the variance coefficients within that range corresponded with large effect size estimates. Additionally, the GIA accounted for 63% to 94% of the predictable variance in the regression models. Conversely, the CHC factors accounted for 3% to 33% of the achievement variance, once the effects of the GIA were controlled for. Whereas the amount of predictable variance accounted for by the factors for the 15 point and 23 point discrepancy subsample was consistent with estimates obtained from analyses of the total sample, the factors for the 30 point discrepancy accounted for 18% to 33% of additional predictable achievement variance. Differences between the variance coefficients obtained for the 30 point difference subsample and those that were obtained from HMR analyses of the total sample was statistically significant across reading, math, and writing outcomes. These results indicate that inter-factor variability has a moderating effect on the predictive validity of the WJ-COG variables.

Effect of SLODR on the Incremental Validity of the WJ-COG

To assess whether Spearman's law of diminishing returns (SLODR) impacted the incremental validity of the broad factors on the WJ-COG, the total study was divided into three subgroups based upon estimated level of general ability, as operationalized through GIA standard scores. The results of the Sequential HMR analyses that were conducted on the below average, average, and above average subgroups can be found in Table 26 (p. 211). Additionally, secondary chi-square goodness of fit tests were conducted that compared differences between the below average and above average group variance coefficients and those obtained from the average/control group. Goodness of fit test results can be found in Table 27 (p. 212).

Average. In the average group (M_{GIA} = 100.47), the GIA accounted for 27% to 33% of dependent variable variance across achievement outcomes on the WJ-ACH. The R^2 coefficients that corresponded with those values are considered to be large effect sizes. The CHC broad factors entered jointly into the second block of the regression equations accounted for 4% (broad mathematics) to 8% (broad reading) of additional achievement variance in the models. The ΔR^2 values that corresponded with those variance increments all represented small effect sizes.

Below average. In the below average group (M_{GIA} = 72.11), the GIA accounted for 24% to 30% of dependent variable variance across achievement outcomes of the WJ-ACH. Whereas those R^2 coefficients are considered to be large effects, they are lower than general factor estimates obtained in other areas of this study. CHC broad factors entered jointly into the second block of the regression equations accounted for 10% (broad mathematics and broad written language) to 11% (broad reading) of additional achievement variance in the models. The ΔR^2 values that corresponded with those variance increments all represented moderate effect sizes. Goodness of fit tests for differences observed between the variance coefficients obtained from the HMR analyses of the below average group and control estimates from the average subsample were statistically significant, for broad mathematics, χ^2 (2) = 9.46, p = .001.

Above average. In the above average group (M_{GIA} = 126.85), the GIA accounted for 11% to 20% of dependent variable variance across achievement outcomes of the WJ-ACH. These R^2 coefficients are considered to be moderate to large effect size estimates and are substantially lower than general factor estimates obtained in other areas of this study. CHC broad factors entered jointly into the second block of the regression equations accounted for 4% to 15% of additional achievement variance in the models. Specifically, in the broad reading model, an additional 57% of predictable variance was accounted for by the factors, whereas in the written language model an additional 56% of predictable variance was accounted for by the factors. When compared to the R^2 coefficients for the GIA, the ΔR^2 values represented increases in prediction by factors of 75% (reading) to 127% (written language). As would be expected from these values, goodness of fit tests for differences observed between the variance coefficients obtained from the HMR analyses of the above average group and control estimates from the average subsample were statistically significant, for broad reading, χ^2 (2) = 11.86, p = .002, broad mathematics, χ^2 (2) = 10.10, p = .001, and broad written language, χ^2 (2) = 21.10, $p < .001$.

Summary. As indicated by SLODR, the GIA accounted for more achievement variance in the below average ability group than in the above average ability group, although the proportions of achievement variance predicted by CHC cognitive abilities were relatively equal across groups. In the case of the above average group, CHC abilities provided for increases in prediction that were equal to or exceeded proportions already accounted for by the general factor. These results provide additional empirical support for the validity of SLODR at the level of prediction with specific groups.

Assessing Relative Achievement Variance Explained by the WJ-GOC Model

Variance estimates were obtained from previous incremental validity studies of other intelligence test measures using the same design parameters as the current study.

The coefficients were then quantitatively synthesized within a fixed effects meta-analytic format for the purposes of calculating grand values, which were then compared to the estimates obtained from the current study via chi-square goodness of fit tests to determine if the variance accounted for by the WJ-COG theoretical model provided for statistically significant increases in achievement variance when compared to estimates obtained from other intelligence test measures.

The R^2 and ΔR^2 coefficients, across reading, math, and writing outcomes for the individual studies that met the inclusion criteria (see Chapter III) are organized by date in a graphic display in Table 28 (p. 213). Across intelligence tests ($N = 6$), the full-scale score R^2 values ranged from .27 to .71 in predicting norm-referenced achievement outcomes. Whereas the effect sizes within that range are all considered to be large, significant variability between the predictive effects of the general factor was demonstrated across tests. Less variability was found in the ΔR^2 values (.00 to .16), which corresponded with the proportion of incremental prediction accounted for by each individual test's factor scores, after controlling for the effects of the full-scale score. Aggregate parameter estimates were obtained for reading, math, and writing, after multiplying each individual effect size estimate by a variance term ($N - 3$) that took into account the sample size for each study. The weighted parameter estimates were than compared to those obtained from the total sample HMR analyses conducted on the WJ-COG in the present study using chi-square goodness of fit tests. The results of the chi-square tests can be found in Table 29 (p. 214).

In the reading model, a $\overline{R^2}$ value of .48 was obtained, 95% CI [.46, .50], indicating that approximately 48% of reading composite variance on norm-referenced tests of achievement was predicted from a synthesis of the full-scale scores of other intelligence test measures. The full-scale effect sizes were significantly heterogeneous (*Q* = 79.66, *p* < .001, I^2 = 91%). As a result, a fixed effects model cannot be assumed at the full-scale level for reading prediction, thus the $\overline{R^2}$statistic can only be interpreted as a descriptive statistic for the current sample. The factor scores contributed approximately 4% of additional predictive variance ($\overline{\Delta R^2}$= .04, 95% CI [.02, .05]). Tests of homogeneity for $\overline{\Delta R^2}$were not significant (*Q* = 1.52, *p* = .98, I^2 = 0%), indicating that a common population effect size was estimated by the studies at the factor level.

In the math model, a $\overline{R^2}$ value of .46 was obtained, 95% CI [.45, .47], indicating that approximately 46% of math composite variance on norm-referenced tests of achievement was predicted from a synthesis of the full-scale scores. The full-scale effect sizes were significantly heterogeneous (*Q* = 65.08, *p* < .001, I^2 = 89%). As a result, a fixed effects model cannot be assumed at the full-scale level for math prediction, thus the $\overline{R^2}$statistic can only be interpreted as a descriptive statistic for the current sample. The factor scores contributed approximately 2% of additional predictive variance ($\overline{\Delta R^2}$= .02, 95% CI [.00, .03]). Tests of homogeneity for $\overline{\Delta R^2}$were not significant (*Q* = 2.36, *p* = .94, I^2 = 0%), indicating that a common population effect size was estimated by the studies at the factor level.

In the writing model, a $\overline{R^2}$ value of .39 was obtained, 95% CI [.37, .39], indicating that approximately 39% of writing composite variance on norm-referenced tests of achievement was predicted by the full-scale scores. The full-scale effect sizes were significantly heterogeneous (*Q* = 34.02, *p* < .001, I^2 = 82%). As a result, a fixed effects model cannot be assumed at the full-scale level for writing prediction, and the $\overline{R^2}$statistic can only be interpreted as a descriptive statistic for the current sample. The factor scores contributed approximately 3% of additional predictive variance ($\overline{\Delta R^2}$= .03, 95% CI [.01, .04]). Tests of homogeneity for $\overline{\Delta R^2}$were not significant (Q = 6.15, p = .41, I^2 = 2%), indicating that a common population effect size was estimated by the studies at the factor level.

When compared to obtained parameter estimates, the WJ-COG model provided for increased amounts of achievement variance at the full-scale and factor score levels, as well as decreased levels of unexplained variance at a descriptive level. Despite these findings, the goodness of fit test results for observed differences in reading, χ^2 (2) = 3.71, p = .16, and writing, χ^2 (2) = 3.33, p = .19, were not statistically significant. The WJ-COG accounted for equal proportions of variance, when compared to parameter estimates, in predicting mathematics outcomes on norm-referenced achievement measures.

Summary. The results of the fixed effects analyses indicate that the amount of achievement variance accounted for by WJ-COG model are consistent with aggregated parameter estimates with similar outcomes variables obtained from other intelligence test measures. Though the variance coefficients deviated slightly from expected values in the reading and writing models, those discrepancies did not exceed predetermined critical levels for statistical significance. Interestingly, coefficients for the math model were virtually equal to expected parameter values. Consistency statistics were not significant when assessing factor level effects, indicating that a common population value was being estimated across the studies included in the analyses, however tests of homogeneity/heterogeneity were significant at the full-scale level, indicating that the obtained aggregate values (i.e., $\overline{R^2}$) can only be interpreted as descriptive statistics for the sample utilized in the current study.

Chapter V: Discussion

The purpose of this study was to evaluate the incremental validity of the Cattell-Horn-Carroll (CHC) broad factors on the Woodcock-Johnson III Tests of Cognitive Abilities (WJ-COG). Previous investigations (e.g., Glutting et al., 1997; Watkins & Glutting, 2000) have shown that similar factor scores on other intelligence tests accounted for negligible portions of achievement variance once the effects of the general factor or full-scale score had been removed. Despite these results, researchers have not examined in depth the incremental validity of assessments that have been designed to measure the constructs within the CHC model of intellectual abilities (McGrew, 2009). Most of the previous predictive validity research has been conducted on different versions of the Wechsler scales (e.g., WISC, WAIS). Although this study was primarily predictive in nature, the results revealed several important findings regarding the efficacy of CHC variables in predicting norm-referenced achievement. These findings offer several important practical and theoretical implications regarding the interpretation of contemporary intelligence test measures and the assessment of specific learning disabilities. This chapter is composed of four major sections: (a) a review of the support found for the research hypotheses outlined in Chapter III, (b) implications for theory and practice, (c) the strengths and limitations of the current study, and (d) conclusions and future research directions.

Support for Research Hypotheses

Hypothesis 1: The CHC factors on the WJ-COG will account for *statistically* significant amounts of incremental achievement prediction on the WJ-ACH above and beyond the effects of the general factor.

The first hypothesis was supported by the results obtained in the current study. After controlling for the effects of the general factor score, the CHC factors accounted for statistically significant amounts of incremental prediction across all seven of the achievement domains assessed on the WJ-ACH. When assessed individually, the contributions of the crystallized ability factor were significant across all seven achievement variables and the fluid reasoning factor provided statistically significant incremental prediction for each achievement domain except math calculation. The auditory processing, visual processing, short-term memory, and processing speed factors provided significant predictive effects for at least two of the IDEA-related achievement clusters. Interestingly, the long-term retrieval factor was only implicated in the reading comprehension regression model. It should be noted that in the reading comprehension model, only the auditory processing factor failed to provide significant incremental predictive effects.

Although these findings are consistent with previous incremental validity research (e.g., Canivez, 2011) regarding the statistical significance of regression results on other intelligence test measures, the practical utility of these findings is questionable. For example, the visual processing factor resulted in incremental prediction estimates that were statistically significant for four out of seven of the regression models, however those estimates only resulted in a 1% total increase in achievement variance accounted for at the practical level. Thus, practitioners who rely on statistical significance as a criterion for evaluating the relative importance of predictor variables run the risk of over-interpreting WJ-COG assessment data.

Hypothesis 2: The CHC factors on the WJ-COG will not account for *clinically* significant amounts of incremental achievement prediction on the WJ-ACH above and beyond the effects of the general factor.

Support for the second research hypothesis was less consistent. For the most part, the CHC factors provided incremental validity estimates consistent with negligible to small effect sizes (e.g., $R^2 < .08$) when predicting achievement outcomes on the WJ-ACH. These coefficients were discrepant from the estimated effects of the general factor, which accounted for large portions of achievement variance across all of the WJ-ACH variables that were assessed in the current study. The relatively low predictive effects for the general factor in the math calculation model was somewhat surprising given previous incremental validity studies have consistently found that the full-scale score on intelligence tests accounted for approximately 50% to 60% of math calculation variance on norm-referenced achievement tests. Additionally, the quantitative reasoning factor has been found to demonstrate strong relationships with *g* at the latent level in studies utilizing structural equation modeling (SEM; see Keith et al., 2006). An explanation for the lower general factor variance coefficient is further confounded by the fact that the content of the math calculation cluster on the WJ-ACH is consistent with similar measures of math calculation on other norm-referenced tests of achievement.

The individual CHC factors failed to account for significant levels of achievement variance in six of the seven regression models. In 31 of the 49 total factor level regression entries, the obtained variance increment for individual factors was equivalent to zero. The fluid reasoning and crystallized ability factors provided for positive incremental prediction in every model except for math calculation whereas, short-term memory and processing speed were implicated in only three of the models. The visual processing and auditory processing factors provided positive incremental prediction in only the math reasoning and oral expression models. It should be noted that most of the incremental variance coefficients associated with these findings corresponded to small effect size estimates. Interestingly, the long-term retrieval factor did not account for any achievement variance beyond effects attributed to the general factor in any of the regression models. These predictive observations complement recent factor analytic studies (e.g., Dombrowski, 2013; Dombrowski & Watkins, 2013) calling into question the reported factor structure in the WJ-COG test manual.

However, an exception was observed in the regression results that were obtained for the oral expression model. In predicting oral expression outcomes, the CHC factors jointly accounted for an increase in predictive variance that was consistent with a large effect size estimate. This is the first study to report incremental effects of this magnitude at the factor level. At the individual level of analysis, the crystallized ability factor alone accounted for almost all of the oral expression variance that was estimated when all of the CHC factors were entered together in the regression equation in the second block. Interestingly, several other factors contributed additional predictive power in increments that exceeded chance factors due to rounding errors. These results indicate that some of the individual-level variance might have been absorbed when the factors were entered jointly into the regression equation. Such suppressor effects have been noted in other HMR designs (Cohen et al., 2003).

There are two potential explanations for the additional variance estimated at the individual factor level. The first was provided by Carroll (1993), who noted that several stratum I abilities had cross-loadings on multiple stratum II factors. Specifically, he noted that narrow visualization ability had significant loadings on both the fluid reasoning and visual processing factors. If a factor scale is composed of multiple latent abilities (e.g., narrow skills) it confounds interpretation of that factor because “different aspects of the scale may have different correlations with external variables” (Reise, Waller, & Comrey, 2000, p. 293).

Another potential explanation for this phenomenon comes from a model of cognition known as “bond theory.” According to Thompson (1916), cognitive tests sample an individual’s brain bonds or neural connections. Thompson’s theory was that the brain might be composed of a large number of these bonds, and that positive manifold was a statistical artifact resulting from overlap in the bonds sampled by similar cognitive tasks. If bond theory is correct, some of the predictive variance accounted for by the unique bonds that are sampled by the qualitatively different CHC broad factors could be lost when force entered together in a regression equation. That lost variance might have been regained when each factor was individually assessed. Despite the intuitive appeal of bond theory, it is important to note that additional variance, beyond that which was attributable to rounding errors, at the individual-level of analysis was only found when predicting oral expression outcomes.

Whereas the implications from these findings indicate that the incremental validity of the CHC factors is established for the prediction of norm-referenced oral expression outcomes, the results are tempered by the potential confound of construct overlap between the predictors and the criterion measure. According to Mather, Wendling, and Woodcock (2001), the oral expression cluster "measures linguistic competency and vocabulary knowledge" (p. 127), and is implicated within the same text as being a measure of crystallized ability. Such predictor-criterion contamination was also observed in a previous study that examined relationships between WJ-COG CHC abilities and math performance across the lifespan. Specifically, Floyd, Evans, and McGrew (2003) noted that overlap between latent constructs (e.g., quantitative reasoning) across tests potentially resulted in spurious increases in predictive power for the fluid reasoning and processing speed factors. Therefore, the predictive effects attributed to the crystallized ability factor should be viewed with caution. Nevertheless, the results of this study indicate that the null hypothesis should be partially rejected with respect to predicting oral expression.

These findings are consistent with previous research on the incremental validity of factor scores on other intelligence test measures with standardization samples and adds to the literature regarding the applied validity of CHC-based factor scores. When considering why the factors generally failed to account for meaningful achievement variance beyond the General Intellectual Ability (GIA) composite, it is important to remember that all factor level scores on intelligence tests contain a mixture of error variance and construct representation from narrow, broad, and general abilities. According to Glutting et al. (2006), "the omnibus IQ measure will contain some variance from the lower order constructs that will take precedence in hierarchical predictive situations" (p. 111). These results indicate that common variance is removed, there is little unique variance left in the CHC measures that is useful for predicting achievement.

Hypothesis 3: The predictive validity of the CHC factors will not be invariant across levels of schooling.

The results from the current study did not support this hypothesis. For the purposes of determining whether or not level of schooling moderated previously obtained incremental validity estimates, the total sample was divided into three subgroups by school level (i.e., primary, intermediate, secondary). When the obtained variance coefficients for each of the grade-level subgroups estimating the predictive effects of the general factor and CHC factors were compared to total sample parameter estimates, the differences were not statistically significant. Across reading, math, and writing outcomes, the predictive effects of the general factor and CHC factors was relatively stable. These results indicate that level of schooling is not a mediating variable for the incremental validity of the CHC factors on the WJ-GOG when utilized to predict norm-referenced achievement outcomes.

The WJ-COG technical manual (McGrew, Schrank, & Woodcock, 2007) provides evidence for discrepant *W*-score (Woodcock & Dahl, 1971) growth patterns among CHC abilities across the lifespan. Namely, crystallized abilities appear to grow linearly over time and are maintained well into adulthood, whereas fluid abilities tend to level off and incrementally decline after the age of 25. However, the *W*-score is an index of proficiency derived from item response theory, which differs from traditional standard scores (e.g., IQ points) that reflect an individual's performance with respect to the normal distribution. Although McGrew and colleagues argued that *W*-score changes reflect legitimate changes in the developmental trajectory of the trait being measured, *W*-score changes do not necessarily correspond to changes in observed standard scores. It is worth noting that in a recent factor-analytic investigation that used the same WJ-COG standardization data from the current study, Tucker-Drob (2009) found inconsistent evidence for systematic changes in CHC abilities across the lifespan. Thus, differential growth patterns may not be significant enough during the school age years to have an impact on the predictive validity of the general factor and/or individual CHC-related abilities.

Despite the non-significant chi-square test results, the predictive effects of the general factor systematically increased across grade levels in the regression models. Although growth was most pronounced in the written language model, the general factor accounted for over 60% of norm-referenced reading performance at the secondary level, the highest variance coefficient found for the GIA in this study. Fascinatingly, the predictive effects of the CHC factors were relatively unitary across all of the models. These findings contradict the prevailing notion (i.e., Jensen, 1998; Mackintosh, 2011) that correlations between IQ and achievement decrease as students become older.

The results obtained in this study provide important information as to the relative stability of the predictive effects of the general factor and broad cognitive abilities across the lifespan (i.e., levels of schooling). Whereas the predictive power of the CHC factors did not fluctuate across levels of schooling, increases in the effects associated with the general factor were noted across reading and writing outcomes. Despite these observations, it should be noted that cognitive effects were assessed at the level of prediction and should not be interpreted as supporting growth in IQ level over time as it relates to predicting achievement outcomes. More research is needed to determine if the findings from the present study generalize to other measures of intelligence and clinical samples.

Hypothesis 4: The predictive validity of the general factor will not be attenuated by significant levels of inter-factor variability on the WJ-COG.

The results of this study supported this hypothesis. Regression analyses conducted with subgroups stratified by level of discrepancy between fluid reasoning (Gf) and crystallized ability (Gc) scores revealed that inter-factor variability was a significant moderating variable in the predictive effects of the CHC factors, in spite of the fact that the predictive effects of the general factor remained strong. Whereas general factor effects were relatively stable and consistent with control group estimates across the discrepancy levels, the incremental validity of the CHC factors improved linearly as Gf-Gc variability increased. In predicting broad reading, the CHC factors provided moderate to large increases in predictive effects across the discrepancy levels. The summed CHC-factor estimate for the 30 point group was larger than any estimate of incremental predictive effects reported in the studies that were reviewed in Chapter I. The differences that were obtained between the group coefficients and expected values from the control group were statistically significant across all achievement outcomes at the 30 point level, and specifically for broad reading at the 15 and 23 point discrepancy levels.

Whereas these results are consistent with previous incremental validity studies (e.g., Freberg, Vandiver, Watkins, & Canivez, 2008; Watkins & Glutting, 2000) with respect to the effects of the general factor, significant departures were noted when assessing the efficacy of the CHC factors in predicting achievement in the presence of significant inter-factor variability. The present study is the first to report such findings. The significance of the factor effects across the broad reading domain is interesting for several reasons. First, predictor-criterion contamination effects are ruled out on the basis of the fact that each broad achievement cluster is composed of a fluency subtest measure, which would theoretically inflate estimates of the predictive power attributable to the processing speed factor across domains. However, incremental prediction was less pervasive for broad math and broad written language. Secondly, previous studies (e.g., Floyd, Evans, & McGrew, 2003; Floyd, McGrew, & Evans, 2008) have implicated higher-order processes such as fluid reasoning and crystallized ability as being consistent predictors of math and writing achievement across the lifespan, whereas relationships between CHC abilities and reading have been less consistent. Despite these implications, small effects for the fluid reasoning and crystallized ability factors were observed in predicting math and writing outcomes for the same 30 point variability subgroup.

Additionally, whereas CHC factor effects were significant in several of the regression models, the stability of the general factor estimates across those conditions indicated that it remains a robust predictor of achievement in the presence of significant levels of Gf-Gc variability. Such a discrepancy occurred in less than 5% of the standardization sample. Thus, the common diagnostic practice of eschewing interpretation of the full-scale IQ score in the presence of significant inter-factor variability (i.e., Kaufman & Lichtenberger, 2006) is not advised when using the WJ-COG to predict achievement. When significant levels of inter-factor variability are observed, these findings indicate that the CHC factors accounted for moderate to large portions of incremental variance across norm-referenced achievement outcomes. In predicting broad achievement at that level of variability, support was found for the simultaneous interpretation of stratum III and stratum II level data. At the individual level of analysis, only the fluid reasoning and crystallized ability factors provided for significant levels of incremental prediction in the broad reading model.

Hypothesis 5: Differential rates of predictive validity will be demonstrated by the CHC factors on the WJ-COG across groups classified by estimated general ability level.

This hypothesis was generally supported, with inconsistencies noted in the domain of broad mathematics. At a descriptive level, general factor variance decreased in predicting outcomes from the below average group to the above average group across all three achievement categories. This finding was consistent with the differentiation hypothesis known as Spearman's law of diminishing returns (SLODR). A second characteristic of SLODR is that group ability factor effects are more prevalent in higher ability groups as a consequence of decreased *g* loadings. In the reading and writing models, CHC factor effects were greater for the above average group than for the the below average group. However, in math, the predictive variance attributable to the CHC factors unexpectedly decreased from the below average group to the high average group. The differences between the above average group and the average group were significant across all of the achievement domains, whereas differences were only significant in math for the below average group.

Previous research has provided evidence for SLODR via latent variable modeling and factor analysis (Reynolds, 2013). Despite these results, Murray et al., (2013) argued that the evidence base for SLODR was largely an artifact of the methods utilized to assess the construct. Specifically, they argued that non-linear *g* loadings could be attributed to manifest skew in the distributions that were assessed. As a result, Murray and colleagues concluded that contemporary assessment techniques (e.g., SEM) were unable to distinguish between test characteristics that had nothing to do with latent ability structures and true SLODR effects. The current study utilized HMR to test for potential SLODR effects at the level of prediction. The results of the current study indicate that SLODR had a consistent moderating effect in predicting reading and writing outcomes, whereas less consistent effects were observed in the math prediction models.

In the above average group, the general factor accounted for a significantly lower portion of achievement variance compared to other estimates obtained within this study. Across the prediction models, the variance coefficients associated with the general factor were consistent, with values that corresponded to moderate effect size estimates. Low general factor values such as these have only been reported in studies using clinical samples (e.g., Nelson & Canivez, 2012). In the below average and average groups, general factor estimates were slightly larger but not within previously estimated ranges (e.g., 40% to 60%). These results indicate that relationships between achievement and the WJ-COG general factor are less stable across subgroups and disaggregated samples than previously thought (i.e., Brody, 1992; Jensen, 1998; Mackintosh, 2011). Furthermore, the substantial increases in unaccounted for variance by the regression equations are large enough to indicate a specification error. Ultimately, more research is needed to determine if these results are an artifact of low base rates or a legitimate challenge to validity estimates obtained with larger sample sizes (Matarrazo & Herman, 1985).

Although small to moderate incremental variance was accounted for by the CHC factors across the regression models, the proportion of additional variance accounted for at the factor level was substantially higher than previous estimates. As an example, in the broad mathematics model for the above average group, the ΔR^2 coefficient for the CHC factors corresponded to a small effect size estimate, however, due to the fact that the general factor accounted for such a low proportion of math achievement in the first block of the regression equation, the CHC factors provided a 31% increase in predictive effects in the model. This example illustrates the danger of interpreting effect size estimates in isolation because the practical validity of factor scores is inextricably tied to the magnitude of effect attributed to the full-scale score.

Overall, the results obtained within this study provide evidence for the effects of SLODR at the level of prediction in the domains of reading and writing. Interestingly, in the above average written language model, the CHC factors accounted for more incremental variance than the general factor, a finding that has never before been reported within the published incremental validity literature. Inconsistent factor level effects were evidenced in the math models. The results obtained within this study add to the growing SLODR literature base, and this is the first study to assess the differentiation hypothesis at the level of prediction.

Hypothesis 6: The predictive utility of the CHC factor structure on the WJ-COG will be consistent with model estimates that have been obtained from HMR analyses of other intelligence tests.

The findings from this study support this hypothesis. Although descriptive differences were observed in variance coefficients obtained from the present incremental validity study of the WJ-COG and estimates from other intelligence tests, those differences were not statistically significant. Interestingly, in predicting norm-referenced math achievement, the coefficients were identical with estimated parameter values. These results indicate that the predictive variance accounted for by the WJ-COG factor model is commensurate with that provided by other intelligence test measures.

Consistency tests were significant across achievement domains at the full-scale score level but not at the factor level. These findings indicate that a fixed effects model cannot be assumed at the full-scale level because the individual full-scale variance coefficients for each test are not estimating the same population parameter. The practical implications of this finding are that the obtained $\overline{R^2}$ can only be interpreted as a summary measure for the studies that were included within this comparison. Different full-scale R^2 coefficients would be expected for studies that utilized different samples with the same test, whereas the $\overline{\Delta R^2}$ value associated with the factor level scores is an appropriate summary statistic regardless of the sample assessed with those particular measures. This is the first study to utilize a meta-analytic framework to synthesize incremental validity outcomes via the R^2 statistic and the results are important for providing a broader context for the findings.

Despite the promise of CHC theory (McGrew, 2009), the CHC model of the WJ-COG appears to account for similar portions of achievement variance when compared to other intelligence test measures. Thus, despite its expanded factor composition and "empirically" derived structure, the predictive validity of the WJ-COG may be commensurate with measures that purport to measure less cognitive factors. Frazier and Youngstrom (2007) suggested that findings such as these likely are the result of a historical increase in the number of factors being measured by contemporary intelligence tests (e.g., WJ-COG), which is an artifact of liberal factor analytic techniques utilized to establish internal validity. In practical terms, Frazier and Youngstrom suggested that newer editions of tests are being developed to measure more cognitive abilities than their predecessors to enhance their "cash validity" (i.e., tests that measure more abilities appear to have more practical utility and thus can be sold for more money than tests that measure a smaller number of abilities). Although many of Frazier and Youngstown's claims cannot be empirically evaluated, recent independent factor analytic investigations (e.g., Dombrowski, 2013; Dombrowski & Watkins, 2013) have raised questions about whether the WJ-COG may be over-factored. The current study complements this work at the level of prediction.

An alternative and more nuanced potential explanation for the mechanism behind these results was provided by Kevin McGrew during an invited address at the inaugural session of the Richard Woodcock Institute for Advancement of Contemporary Cognitive Assessment. According to McGrew (2012), the predictive validity of cognitive tests has remained relatively unchanged over the last half century because of an inherent limitation in assessing cognitive-achievement relationships through a conventional linear framework, and, as such, conventional ability measures are incapable of capturing the dynamic fundamentals that underlie most cognitive-achievement interactions. Therefore, new theoretical advances (e.g., CHC), and the inclusion of additional broad ability factors on contemporary IQ measures have done little to provide enhanced levels of prediction because they assume linear compartmentalized relationships between specific cognitive abilities and achievement.

Over the past 5 years, McGrew and colleagues have conducted an ambitious research program that examines cognitive-ability relationships using multi-dimensional scaling to assess for trait complexes that best predict achievement. Although in its infancy, this work has led McGrew to conclude that the most optimal predictors of achievement are multi-faceted cognitive measures that combine functionally similar aptitude and scholastic abilities to maximize cognitive load. As an example, a measure that requires the simultaneous use of working memory and processing speed abilities requires more cognitive load than a pure measure of processing speed. Whereas the CHC framework continues to be a useful taxonomy for classifying and organizing cognitive abilities, the construction of assessments via broad CHC factors fails to take into consideration the role of cognitive complexity in the design process (McGrew, 2012). As a result, the search for factorially pure broad CHC measures has diluted the predictive validity of test batteries such as the WJ-COG.

Although beyond the scope of the current study, McGrew's fusion of developmental science, scholastic abilities, and cognitive aptitudes raises interesting questions about optimal cognitive test design and the evolution of CHC theory. This work also has culminated in an interesting treatise, which has the potential to shed some light on the large portion of variance that has historically been left unexplained by cognitive-achievement prediction models (see McGrew, 2008). In the current study, approximately half of the criterion variance in the regression models was left unexplained.

Summary of Results

As a whole, these results provide an interesting look at the predictive relationships between cognitive-achievement variables on the WJ-III. Overall, three of the research hypotheses were supported; two were partially supported, and one was not supported by the data. Whereas the purpose of this study was to assess the incremental validity of the WJ-COG, the results illustrate that incremental validity may not be a unitary construct and can vary depending on the particular conditions and samples that are assessed. Nevertheless, the results obtained within this study help to clarify procedures for interpreting the WJ-COG.

In all of the areas that were assessed in the current study, the GIA accounted for the largest and most consistent amounts of achievement variance across the regression models. Most of the variance coefficients associated with the general factor corresponded to large effect size estimates. The general factor remained a robust predictor of achievement across significant levels of inter-factor variability and levels of schooling. It is important to note that not all general factor effects were as robust. The coefficients obtained for many of the secondary research areas were commensurate with values found in previous studies that utilized clinical samples. The results indicate that SLODR may have an attenuating effect on the predictive validity of the general factor, resulting in factor scores accounting for more achievement variance than the general factor in one of the regression models.

Evidence was found for the incremental validity of the CHC factors in predicting oral expression with the total standardization sample, reading and writing outcomes in the presence of significant inter-factor variability, and when predicting achievement outcomes after accounting for the effects of SLODR. Effects at the individual factor level were small for most of the CHC factors, although the crystallized ability and fluid reasoning factors provided moderate predictive effects in several of the regression models.

Implications for Practice

Several implications for practice can be drawn from the results obtained from the current study. First, although some evidence was found for the incremental validity of the CHC factor/cluster scores on the WJ-COG, these results do not support the statement made in the WJ-COG technical manual that the CHC factor scores “provide the primary basis for interpretation” (Mather & Woodcock, 2011b, p. 11). In contrast, the current study indicates that the GIA should be given the greatest interpretive weight as it accounted for the largest amount of variance across achievement outcomes and models of assessment. The GIA consistently accounted for greater portions of achievement variance than that accounted for by the CHC factor scores. Therefore, practitioners who forego interpreting the GIA in favor of the factor scores do so at the expense of eliminating the most reliable and valid construct available on the WJ-COG.

Second, these results provide more support for secondary-level clinical interpretation of the factor scores than the findings from previous incremental validity studies (e.g., Glutting et al., 2006). Whereas the combined effects of the fluid reasoning, crystallized ability, auditory processing, visual processing, short-term memory, long-term retrieval, and processing speed factors accounted for similar portions of incremental achievement prediction for most academic outcomes, the CHC factor scores predicted meaningful oral expression achievement beyond the GIA. However, at present there exists no interpretive strategy for disentangling the clinical implications of joint factor effects. As a result, the practical utility of these observations is not known.

Furthermore, individual factors failed to account for meaningful predictive variance beyond the GIA in all of the regression models except for oral expression and broad reading in the presence of 30 point Gf-Gc variability. As previously noted, the Gf and Gc factors accounted for moderate portions of variance beyond the effects of the general factor in these models. Leaving potential predictor-criterion contamination effects aside, interpretation of the Gf and Gc factors is confounded by additional theoretical constraints. Namely, it has yet to be established whether these factors are appropriately identified as stratum II abilities, as they have been implicated as proxies for the *g*-factor (see Cattell, 1963). It is interesting to note that Cattell originally argued that *g* was better represented as a bifurcated ability, via Gf and Gc, with additional broad abilities subordinate to those factors. His "investment theory" (Cattell, 1987) later provided an explanation as to how these two factors interacted throughout the lifespan to produce intelligent behavior. Only recently, as a result of the development of CHC theory, have Gf and Gc been reassigned as stratum II level abilities commensurate with lower level cognitive processing factors. Thus, practitioners should be cautious of interpreting these factors in the same manner as they would more discrete factors such as auditory processing or short-term memory.

Third, empirical support for the practice of eschewing interpretation of the full-scale IQ score in the presence of significant levels of inter-factor variability was not found for the WJ-COG. Across multiple levels of Gf-Gc variability, the GIA accounted for large effects in predicting reading, writing, and math outcomes. Variability levels that would be expected in 5% or less of the population did not have an attenuating effect on the predictive validity of the GIA. Although, the CHC factors provided moderate to large incremental effects in several of the models, the results from the current study indicate that primary interpretation should be at the general factor level.

The results from the current study also have implications that are relevant for evaluating the practical validity of models (e.g., Flanagan, Alfonso, & Mascolo, 2011) that require practitioners to evaluate relationships between an individual's profile of cognitive strengths and weaknesses and achievement markers (PSW) for the purposes of specific learning disability (SLD) identification. In the CHC-based PSW model advocated by Flanagan and colleagues, SLD is psychometrically operationalized as a link between a normative cognitive deficit in a CHC/neuropsychological domain and a concomitant deficit in a relevant area of achievement, with remaining psychoeducational abilities falling within expected ranges in this PSW model. Although a full-scale IQ score can be utilized as evidence to demonstrate a pattern of cognitive behavior not related to the deficit area of concern, primary interpretation is to occur at the stratum I and stratum II levels of the CHC model. According to Flanagan et al., "it is this very notion that makes it necessary to draw upon cognitive and neuropsychological theory and research to inform operational definitions of SLD and increase the reliability and validity of the SLD identification process" (2011, p. 250). However, the results from the current study indicate that practitioners who interpret CHC factor scores on the WJ-COG, without accounting for the effects of the GIA, risk overestimating the predictive effects of various CHC abilities. The HMR analyses with the total sample indicate that the CHC factor scores contain a substantial amount of common variance which reflects the effects of *g*. Once this common variance is controlled for, the remaining unique variance accounted for negligible effects across most of the regression models. When evaluating

PSW approaches to SLD identification that require interpretation at the factor or broad ability level, practitioners are encouraged to consider the following admonition from Carroll (1976):

> Nearly all cognitive tasks are complex, in the sense that they involve different kinds of memories and control processes…It may be impossible, in principle, to identify 'pure' factors of individual differences…The often-noted observation that all psychometric tests in the cognitive domain are more or less positively correlated probably reflects the multifaceted nature of the tasks sampled on those tests. (p. 52)

Whereas it may be possible for practitioners to account for *g* effects when interpreting primarily at the factor level, contemporary PSW models have yet to provide a mechanism for doing so.

In response to the growing literature base calling into question the validity of factor and subtest scores on contemporary intelligence test batteries, Glutting and colleagues (2003) raised the question as to whether multi-factored intelligence test batteries were worth their cost in additional administration time and capital resources. Although the answer to that question is beyond the score of the present study, the results obtained here provide potentially valuable information regarding the practical utility of the WJ-COG. Researchers (e.g., Canivez, 2013b; Glutting, Watkins, & Youngstrom, 2003) have argued that in the absence of incremental validity, practitioners should disregard multi-factored IQ tests in favor of brief instruments that are more cost effective. As this study has demonstrated, incremental validity may not be a unitary construct and can fluctuate across a variety of assessment models and conditions. Although the interpretive guidelines for the R^2 statistic provided by Cohen (1988) are useful for interpreting HMR validity coefficients, examples were observed within this study in which the CHC factors provided for meaningful portions of achievement variance beyond the GIA, even though the incremental coefficients could be interpreted as small effect size estimates. In appraising the incremental validity of factor scores, practitioners are encouraged to consider the proportion of variance accounted for at the factor level in the context of the predictive power of the general factor estimate.

Finally, it is important for practitioners to keep in the mind that although the GIA accounted for large achievement effects in most of the samples, large amounts of achievement variance were still left unaccounted for. Even with additional interpretation of broad cognitive abilities, 30% to 50% of achievement outcomes were left unexplained by the prediction models. In sum, general ability is an important factor to account for when assessing children and adolescents with academic concerns, but it is certainly not the only factor that should be appraised.

Strengths of the Current Study

This study contained several strengths. First, the study utilized a large sample size with data from the WJ-COG, which has been designed to measure cognitive abilities from a CHC perspective. The WJ-COG is the only contemporary measure of intellectual functioning that purports to assess cognitive abilities exclusively from a CHC perspective. Additionally, the WJ-COG is frequently utilized in research settings to evaluate the validity of CHC theory as well as the structure and functioning of CHC-related abilities.

The use of HMR to partition variance between the predictor variables also was a strength. By controlling for the effects of the GIA, the unique contributions provided by the broad abilities could be appraised after removing the common or *g* variance. Many studies using other methods (e.g., standard multiple regression, SEM) to assess cognitive-achievement relationships have failed to account for general factor effects that are present in all factor level measures.

Methodologically, this study provided for a comprehensive assessment of incremental validity that included several areas that had previously not been addressed within the literature. Specifically, assessing for the effects of level of schooling as well as accounting for the potential effects of SLODR tested the degree to which the predictive validity of the CHC factors on the WJ-COG was invariant across different samples and conditions.

Limitations of the Current Study

Despite these strengths, this study is not without limitations that should be considered when interpreting the results. First, it is important to remember that this study was designed to be predictive in nature, which limits the explanatory inferences that can be drawn from the data. Whereas some methodologists (e.g., Reynolds & Keith, 2013) have preferred to make a clear distinction between the two epistemologies, predictive studies have implications for explanatory research, though they cannot be utilized alone for explanatory purposes (Hempel, 1965). Whereas the present study provides important practical information regarding predictive relationships between CHC abilities on the WJ-COG and achievement that have implications for interpretation of the WJ-COG in clinical practice by assessment professionals, these results cannot be utilized to make inferences regarding cognitive-achievement relationships at the latent level. Contemporary validity models are fragmented, which requires the integration of research using a variety of statistical methods and samples (Decker, 2013). Given the stark differences that have been obtained between validity studies using HMR and those utilizing other methods (e.g., SEM), researchers must consider the potential impact of method variance when interpreting their research results.

Second, the sequential assessment of individual predictor variables in separate regression models is akin to conducting multiple orthogonal contrasts and can result in an underestimated Type I (incorrectly rejecting a false null hypothesis) error rate. The current study utilized a consistent alpha level for evaluating the statistical significance of test results. The use of appropriate multivariate statistical techniques (e.g., multiple regression) allows for a researcher to assume a constant alpha level across multiple independent variables (a.k.a., family-wise error rate). However, power is lost when IVs are simultaneously assessed in separate regression models. As a rule of thumb, power is lost exponentially with each successive comparison (e.g., regression model) as a protection against finding spurious results when assessing a large number of variables (Burt-Vesey, Vesey, Stroter, & Middleton, 2011). Thus, with each dependent variable analyzed by seven different regression models (one per each CHC factor), there were a total of seven tests on the null hypothesis for the obtained R^2 coefficients associated with each predictor variable. Because these seven tests are not independent, exact investigation-wise (correction for multiple comparisons) error rates cannot be calculated. For the sake of approximation, the error term for the current study was closer to $(1 -$

.95)[7]. As a result of this limitation, caution should be utilized in interpreting the results of the ANOVA tests on the obtained R^2 coefficients.

Third, although the sample sizes for many of the regression models were adequate, a potential threat to statistical conclusion validity was noted as a result of the sample sizes that were included in the 30 point Gf-Gc difference regression models. As was previously discussed in Chapter III, small sample sizes were expected for these groups based upon prevalence rates for that level of discrepancy. As a result, CHC factors were entered one at a time into the second block of the regression equation to reduce the number of predictor variables, thereby increasing power to acceptable levels for data analysis (e.g., $\geq .80$). Unfortunately, as a consequence of using listwise deletion procedures to treat missing data, the obtained sample sizes for the above subgroups fell below expected levels. Post-hoc sensitivity analysis indicated that the sample sizes in the 30 point difference models may not have been adequate to detect small to moderate incremental differences for the individual CHC factors. Based upon a power level of .80 and an alpha of .05, the smallest ΔR^2 value that could be detected reliably was .12. According to Cohen (1988), a coefficient of this magnitude represents a moderate effect size estimate. Although joint effects were estimated to be significant in all of the 30 point models, seven of the incremental variance coefficients for the individual CHC factors in those models were equal to zero.

Another limitation was noted with respect to the composition of the SLODR groups, which were constructed on the basis of observed GIA scores. By limiting the degree to which GIA scores could vary, the range of possible predictor values was artificially restricted. Restriction of range poses a threat to reliability of measurement due to the fact that it can potentially attenuate a bivariate correlation, although range restriction is less of a factor for regression coefficients under the conditions of homoscedasticity and linearity (Cohen et al., 2003). Nevertheless, it is possible that some of the negative predictive relationships that were observed for the CHC variables may have been an artifact of range restriction. Due to the nature of SLODR, the threat of range restriction is a necessary condition for assessing this area of cognitive functioning.

Next, collinearity diagnostics indicated that the crystallized ability, long-term retrieval, short-term memory, and processing speed factors were correlated at an excessively high level with the GIA across the reading, mathematics, and writing regression models. This is the result of the fact that subtests are linearly combined to produce factor and composite scores on the WJ-COG. Although collinearity does not have an impact on variance coefficients (i.e., R^2), Hale and colleagues (2001) have argued that collinearity renders HMR an invalid method to assess incremental validity. According to Canivez (2013a), this redundancy is precisely the problem that practitioners must confront when simultaneously interpreting full-scale and factor scores when conducting assessments of individual cognitive functioning. Although a potential solution exists in simply eliminating interpretation of the GIA and other full-scale IQ composites, doing so is at odds with existing psychological theory. Whether one believes in the utility of the general factor, all second-order (factors) and first-order (subtests) cognitive measures contain varying degrees of common variance that must be accounted for when interpreting tests at those levels.

Finally, there are limitations in using norm-referenced achievement measures to make inferences regarding scholastic performance for individuals. In a study designed to measure the content overlap between norm-referenced measures of reading and the local school curriculum, Good and Salvia (1988) found significant differences in the match between local reading curriculum and the content assessed in the reading measures. Although an update of this research is needed to assess the efficacy of modern instruments, the results indicated that norm-referenced achievement tests have the potential to under-predict actual school performance. Therefore, measures such as the WJ-ACH are best interpreted as screening instruments and are not recommended for diagnostic purposes (Shapiro, 2011).

Implications for Future Research

The results of the current study raise a number of questions to guide future research. First, it is necessary to examine whether these results generalize to clinical samples. The use of standardization data is useful for examining cognitive-achievement relationships in a broader context; however the results from these data may not generalize to relevant decision-making contexts (e.g., school or clinic settings). According to Decker (2013), this is the result of the fact "that validity research is often conducted under controlled conditions, which may not always be relevant for the contextual issues in applied practice" (p. 39). Incremental validity studies using clinical samples have consistently found much lower general factor effects in addition to greater proportions of total variance accounted for by factor level scores. Additional research is needed to determine if similar results are observed in clinical samples with the WJ-COG.

Additionally, research by McGrew and colleagues (e.g., McGrew, 2012; Schneider & McGrew, 2012) has called into question the predictive validity of test designs that utilize strict partitioning of broad cognitive factors. In a recent paper, McGrew (2012) argued that narrow CHC abilities are better predictors of achievement not because of direct narrow ability relationships, but due to the fact that stratum I measures tend to be more "cognitively complex" than broad measures (i.e., assess multiple cognitive abilities simultaneously). Interestingly, the WJ-COG organizes subtests from multiple broad areas into several clinical clusters. As an example, a *cognitive efficiency* cluster is formed by combining the narrow abilities of perceptual speed (P; processing speed) and working memory capacity (WM; short-term memory). Additional research is needed to determine the clinical utility of these cluster scores (Schrank, Miller, Wendling, & Woodcock, 2010).

Finally, variance accounted for by the GIA across many of the subgroups raises questions about whether the conventional wisdom that "IQ accounts for approximately 50% of achievement" (Brody, 1992; Jensen, 1998; Mackintosh, 2011) remains accurate across assessment models and clinical samples. The results from this study indicate that the GIA accounted for lower portions of achievement when accounting for inter-factor variability and SLODR. These estimates are commensurate with estimates of the predictive effects of full-scale composites on other intelligence tests with clinical samples (e.g., Nelson & Canivez, 2012; Nelson, Canivez, & Watkins, 2013). Well-designed research that assesses the validity of interpretive strategies of cognitive assessments with referred samples is especially needed to advance the science of cognitive assessment.

Conclusions

The results of this study indicate that the WJ-COG GIA was consistently the strongest predictor of academics on the WJ-ACH and should be given the greatest interpretive weight by practitioners. In limited circumstances, the CHC factors accounted for meaningful portions of variance beyond the GIA. Whereas no individual CHC factors were consistently implicated in the achievement models, the results of this study are potentially important for several reasons. First, this is the only study to assess the predictive validity of CHC-based factor scores via measures designed specifically to assess CHC-related constructs. Second, this is the first study utilizing standardization data to find that factor level scores on a contemporary measure of intelligence accounted for meaningful predictive variance beyond the effects of the full-scale score using an HMR design. Next, the results of this study demonstrate that predictive effects of the GIA may be moderated by conditions such as SLODR.

References

Ackerman, P. L. (1996). A theory of adult intellectual development: Process, personality, interests, and knowledge. *Intelligence, 22,* 227-257. doi:10.1016/S0160-\ 2896(96)90016-1

Alfonso, V. C., Flanagan, D. P., & Radwin, S. (2005). The impact of Cattell-Horn-Carroll theory on test development and interpretation of cognitive and academic abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment, Theories, tests, and issues* (2nd ed., pp. 185-202). New York: Guilford Press.

Alfonso, V. C., Oakland, T. D., LaRocca, R., & Spanakos, A. (2000). The course on individual cognitive assessment. *School Psychology Review, 29,* 52-64. Retrieved from http://www.nasponline.org

Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA: Sage.

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

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

Becker, K. A. (2003). *History of the Stanford-Binet intelligence scales: Content and psychometrics.* (Stanford-Binet Intelligence Scales, Fifth Edition Assessment Service Bulletin No. 1). Itasca, IL: Riverside.

Belsey, D. A. (1991). *Conditioning diagnostics: Collinearity and weak data in regression*. New York: Wiley.

Berry, W. D. (1993). *Understand regression assumptions*. Newbury Park, CA: Sage.

Berry, W. D., & Feldman, S. (1985). *Multiple regression in practice*. Thousand Oaks, CA: Sage.

Boring, E. G. (1929). *A history of experimental psychology*. New York: Century.

Bowman, M. L. (1989). Testing individual differences in ancient China. *American Psychologist, 44,* 576-578. doi: 10.1037%2f%2f0003-066X.44.3.576.b

Brockmeier, L.L., Kromrey, J. D., & Hines, C. V. (1998). Systematically missing data and multiple regression analysis: An empirical comparison of deletion and imputation techniques. *Multiple Linear Regression Viewpoints, 25 (1),* 20-39. Retrieved from: http://mlrv.ua.edu/1998/VOL25_N1_A5.pdf

Brody, N. (1992). *Intelligence* (2nd ed.). San Diego, CA: Academic Press.

Brody, N. (1997). Intelligence, schooling, and society. *American Psychologist, 52,* 1046-1050. doi: 10.1037/0003-066X.52.10.1046

Bulmer, M. G. (1979). *Principles of statistics*. New York: Dover Publications.

Burns, R. B. (1994). Surveying the cognitive terrain. Reviewed work(s): *Human Cognitive Abilities: A Survey of Factor Analytic Studies* by John B. Carroll. *Educational Researcher, 23 (3),* 35-37. doi:10.3102/0013189X023003035

Burt-Vesey, W. M., Vesey, J.V., Stroter, A., & Middleton, K. (2011). Multiple linear regression: A return to basics in educational research. *Multiple Linear Regression Viewpoints, 37(2),* 14-22. Retrieved from http://mlrv.ua.edu/2011/vol37_2/ Veseyetal_37_2_Final_2.pdf

Canivez, G. L. (2008). Orthogonal higher-order factor structure of the Stanford-Binet Intelligence Scales for children and adolescents. *School Psychology Quarterly,* *23,* 533-541. doi: 10.1037/a0012884

Canivez, G. L. (2011, February). *Incremental predictive validity of Cognitive Assessment System PASS scores*. Paper presented at the 2011 Annual Convention of the National Association of School Psychologists, San Francisco, CA.

Canivez, G. L. (2013a). Incremental criterion validity of WAIS-IV factor index scores: Relationships with WIAT-II and WIAT-III subtest and composite scores. *Psychological Assessment, 25,* 484-495. doi: 10.1037/a0032092

Canivez, G. L. (2013b). Psychometric versus actuarial interpretation of intelligence and related aptitude batteries. In D. H. Saklofske, C. R. Reynolds, & V. L. Schwean (Eds.). *The oxford handbook of child psychological assessment* (pp. 84-112). New York: Oxford University Press.

Canivez, G. L., & Watkins, M. W. (1998). Long term stability of the Wechsler Intelligence Scale for Children-Third Edition. *Psychological Assessment, 10,* 285-291. doi: 10.1037/1040-3590.10.3.285

Canivez, G. L., & Watkins, M. W. (2010). Investigation of the factor structure of the Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV): Exploratory and higher-order analyses. *Psychological Assessment, 22*, 827-836. doi:10.1037/a0020429

Carroll, J. B. (1976). Psychometric tests as cognitive tasks: A new structure of intellect. In L. B. Resnick (Ed.), *The nature of intelligence* (pp. 27-56). Hillsdale, NJ: Erlbaum.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analytic studies*. New

York: Cambridge University Press.

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 Press.

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

Cattell, R. B. (1944). Psychological measurement: Normative, ispsative, and interactive. *Psychological Review, 51*, 292-303. doi: 10.1037/h0057299

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54,*1-22. doi: 10.1037/h0046743

Cattell, R. B. (1987). *Intelligence: Its structure, growth, and action*. New York: Elsevier Science.

Cizek, G. J. (2003). Review of the Woodcock-Johnson III. In B. S. Plake & J. C. Impara (Eds.), *The fifteenth mental measurements yearbook* (pp. 10240-1024). Lincoln, NE: Buros Institute of Mental Measurements.

Cohen, J. (1959). The factorial structure of the WISC at ages 7-6, 10-6, and 13-6. *Journal of Consulting Psychology, 23,* 285-299. doi: 10.1037/h0043898

Cohen, J. (1988). *Statistical power analysis for the behavior sciences* (2nd ed.). New York: Psychology Press.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple* *regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Crocker, L. C., & Algina, J. (2008). *An introduction to classical and modern test theory*. Mason, OH: Cengage Learning.

Dana, J., & Dawes, R. (2007). Comment on Fiorello et al. Interpreting Intelligence Test Results for Children with Disabilities: Is Global Intelligence Relevant? *Applied Neuropsychology, 14,* 21-25. doi: 10.1080/09084280701280379

Daniel, M. H. (1997). Intelligence testing: Status and trends. *American Psychologist, 52,* 1038-104. doi: 10.1037/0003-066X.52.10.1038

Darlington, R. B. (1990). *Regression and linear models*. New York: McGraw-Hill.

Davis, F. B. (1959). Interpretation of differences among averages and individual test scores. *Journal of Educational Psychology, 50,* 162-170. doi: 10.1037/h0044024

Deary, I. J., Egan, V., Gibson, G. J., Austin, E., Brand, C. R., & Kellaghan, T. (1996). Intelligence and the differentiation hypothesis. *Intelligence, 23,* 105-132. doi: 10.1016/S0160-2896(96)90008-2

Decker, S. L. (2013). Testing: The measurement and assessment link. In D. H. Saklofske, C. R. Reynolds, & V. L. Schwean (Eds.). *The oxford handbook of child psychological assessment* (pp. 30-47). New York: Oxford University Press.

DeGroot, A. D. (1969). *Methodology: Foundations of inference and research in the behavioral sciences*. The Hague: Mouton.

Detterman, D. K. (2002). General intelligence: Cognitive and biological explanations. In R. J. Sternberg & E. L. Grigorenko, (Eds.), *The general factor of intelligence: How general is it?* (pp. 223-243). Mahwah, NJ: Lawrence Erlbaum Associates.

Detterman, D. K., & Daniel, M. H. (1989). Correlations of mental tests with each other and with cognitive variables are the highest for low IQ groups. *Intelligence, 13,* 349-310. doi: 1016/S0160-2896(89)80007-8

DiStefano, C., & Dombrowski, S. C. (2006). Investigating the theoretical structure of the Stanford-Binet –Fifth Edition. *Journal of Psychoeducational Assessment, 24,* 123-136. doi: 10.1177/0734282905285244

Dombrowski, S. C. (2013). Investigating the structure of the WJ-III Cognitive at school age. *School Psychology Quarterly, 28,* 154-169. doi: 10.1037/spq0000010

Dombrowski, S. C., & Watkins, M. W. (2013). Exploratory and higher-order factor analysis of the WJ-III full-test battery: A school-aged analysis. *Psychological assessment, 25,* 442-455. doi: 10.1037/a0031335

Elliott, C. D. (1990). The nature and structure of children's abilities: Evidence from the Differential Ability Scales. *Journal of Psychoeducational Assessment, 8,* 376-390. doi: 10.1177/073428299000800313

Elliott, C. D. (2007). *Differential Ability Scales-Second Edition*. San Antonio, TX: Harcourt Assessment.

Elliott, C. D., Hale, J. B., Fiorello, C. A., Dorvil, C., & Moldovan, J. (2010). Differential Ability Scales-II prediction of reading performance: Global scores are not enough. *Psychology in the Schools, 47,* 698-720. doi: 10.1002/pits.20499

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.

Evans, J. J., Floyd, R. G., McGrew, K. S., & Leforgee, M. H. (2002). The relations between measures of Cattell-Horn-Carroll (CHC) cognitive abilities and reading achievement during childhood and adolescence. *School Psychology Review, 31,* 246-262. Retrieved from http://www.nasponline.org

Fagan, T. K., & Wise, P. S. (2007). *School psychology: Past, present, and future* (3rd ed.). Bethesda, MD: National Association of School Psychologists.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. (2009). Statistical power analysis using G*Power 3.1: Tests for correlation and regression analysis. *Behavioral Research Methods, 41,* 1149-1160. doi: 10.3758/BRM.41.4.1149

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice, 40,* 532-538. doi: 10.1037/a0015808

Fiorello, C. A., Hale, J. B., Holdnack, J. A., Kavanagh, J. A., Terrell, J., & Long, L. (2007). Interpreting intelligence test results for children with disabilities: Is global Intelligence relevant? *Applied Neuropsychology, 21.* 2-12. doi: 10.1080/09084280701280338

Fiorello, C. A., Hale, J. B., McGrath, M., Ryan, K., & Quinn, S. (2002). IQ interpretation for children with flat and variable test profiles. *Learning and Individual Differences, 13,* 115-125. doi: /10.1016/S1041-6080(02)00075-4

Flanagan, D. P. (2003). Use of the Woodcock-Johnson III within the context of a modern operational definition of learning disability. In F. A. Schrank & D. P. Flanagan (Eds.), *WJ III clinical use and interpretation: Scientist-practitioner perspectives* (pp. 199-227). San Diego, CA: Academic Press.

Flanagan, D. P., Alfonso, V. C., & Mascolo, J. T. (2011). A CHC-based operational definition of SLD: Integrating multiple data sources and multiple data-gathering methods. In D. P. Flanagan & V. C. Alfonso (Eds.), *Essentials of specific learning* *disability identification* (pp. 233-298). Hoboken, NJ: John Wiley.

Flanagan, D. P., Alfonso, V. C., & Ortiz, S. O. (2012). The cross-battery assessment approach. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 459-483). New York: Guilford Press.

Flanagan, D. P., & Kaufman, A. S. (2004). *Essentials of WISC-IV assessment*. Hoboken, NJ: John Wiley.

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

Flanagan, D. P., Ortiz, S. O., Alfonso, V. C., & Dynda, A. (2006). Integration of response-to-intervention and norm-referenced tests in learning disability identification: Learning from the Tower of Babel. *Psychology in the Schools, 43*, 807-825. doi: 10.1002/pits.20190

Fletcher, J. M., Lyon, G. R., Fuchs, L. S., & Barnes, M. (2007). *Learning disabilities: From identification to intervention*. New York: Guilford Press.

Fletcher-Janzen, E. (2009). Intelligent testing: Bridging the gap between classical and romantic science in assessment. In J. C. Kaufman (Ed.), *Intelligent testing: Integrating psychological theory and clinical practice* (pp. 15-29). New York: Cambridge University Press.

Floyd, R. G., Bergeron, R., McCormack, A. C., Anderson, J. L., & Hargrove-Owens, G. L. (2005). Are Cattell–Horn–Carroll (CHC) broad ability composite scores exchangeable across batteries? *School Psychology Review, 34,* 386-414. Retrieved from http://www.nasponline.org

Floyd, R. G., Clark, M. H., & Shadish, W. R. (2008). The exchangeability of intelligent quotients: Implications for professional psychology. *Professional Psychology: Research and Practice, 39,* 414-423. doi: 10.1037/0735-7028.39.4.414

Floyd, R. G., Evans, J. J., & McGrew, K. S. (2003). Relations between measures of Cattell-Horn-Carroll (CHC) cognitive abilities and mathematics achievement across the school-age years. *Psychology in the Schools, 40,* 155-171. doi: 10.1002/pits.10083

Floyd, R. G., Keith, T. Z., Taub, G. E., & McGrew, K. S., (2007). Cattell-Horn-Carroll cognitive abilities and their effects on reading decoding skills: *g* has indirect effects, more specific abilities have direct effects. *School Psychology Quarterly, 22,* 200-233. doi:10.1037/1045-3830.22.2.200

Floyd, R. G., McGrew, K. S., & Evans, J. J. (2008). The relative contributions of the Cattell-Horn-Carroll cognitive abilities in explaining writing achievement during childhood and adolescence. *Psychology in the Schools, 45,* 132-144. doi: 10.1002/pits.20284

Floyd, R. G., Shands, E. I., Rafael, F. A., Bergeron, R., & McGrew, K. S. (2009). The dependability of general-factor loadings: The effects of factor-extraction methods, test battery composition, test battery size, and their interactions. *Intelligence, 37,* 453-465. doi: 10.1016/j.intell.2009.05.003

Fox, J. (1991). *Regression diagnostics*. Newbury Park, CA: Sage.

Frazier, T. W., & Youngstrom, E. A. (2007). Historical increase in the number of factors measured by commercial tests of cognitive ability: Are we overfactoring? *Intelligence, 35,* 169-182. doi: 10.1016/j.intell.2006.07.002

Freberg, M., Vandiver, B. J., Watkins, M. W., & Canivez, G. L. (2008). Significant factor score variability and the validity of the WISC-III full scale IQ in predicting later academic achievement. *Applied Neuropsychology, 15,* 131-139. doi: 10.1080/09084280802084010

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

Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in psychology and education* (3rd ed.). Boston: Allyn and Bacon.

Glutting, J. J., Watkins, M. W., Konold, T. R., & McDermott, P. A. (2006). Distinctions without a difference: The utility of observed versus latent factors from the WISC-IV in estimating reading and math achievement on the WIAT-II. *Journal of Special Education, 40,* 103-114. doi: 10.1177/00224669060400020101

Glutting, J. J., Watkins, M. W., & Youngstrom, E. A. (2003). Multi-factored and cross-battery assessments: Are they worth the effort? In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological & educational assessment of children: Intelligence, aptitude, and achievement*. (2nd ed., pp. 343-376). New York: Guilford Press.

Glutting, J. J., Youngstrom, E. A., Ward, T., Ward, S., & Hale, R. L. (1997). Incremental efficacy of the WISC-III factor scores in predicting achievement: What do they tell us? *Psychological Assessment, 9,* 295-301. doi: 10.1037/1040-3590.9.3.295

Good, R. H. III, & Salvia, J. (1988). Curriculum bias in published, norm-referenced reading tests: Demonstrable effects. *School Psychology Review, 17,* 51-60.

Retrieved from http://www.nasponline.org

Gordon, R. A. (1968). Issues in multiple regression. *American Journal of Sociology. 73,* 592-616. doi: 10.1086/224533

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

Gottfredson, L.S. (2005a). Implications of cognitive differences for schooling within diverse societies. In C. L. Frisby & C. R. Reynolds (Eds.), *Comprehensive handbook of multicultural school psychology* (pp. 517-554). Hoboken, NJ: John Wiley.

Gottfredson, L. S. (2005b). Suppressing intelligence research: Hurting those we intend to help. In R. H. Wright & N. A. Cummings (Eds.), *Destructive trends in mental health: The well-intentioned path to harm* (pp. 155-186). New York: Taylor and Francis.

Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach*. New York: Psychology Press.

Groth-Marnat, G. (2009). *Handbook of psychological assessment* (5th ed.). Hoboken, NJ: John Wiley.

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

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

Guttman, L., & Levy, S. (1991). Two structural laws for intelligence tests. *Intelligence, 15,* 79-103. doi: 10.1016/0160-2896(91)90023-7

Hale, J. B., & Fiorello, C. A. (2001). Beyond the academic rhetoric of *g*: Intelligence testing guidelines for practitioners. *The School Psychologist, 55(4),* 113-139. Retrieved from http://www.apadivisions.org/division-16/publications/newsletters/school-psychologist

Hale, J. B., & Fiorello, C. A. (2004). *School neuropsychology: A practitioner's handbook*. New York: Guilford.

Hale, J. B., Fiorello, C. A., Bertin, M., & Shermin, R. (2003). Predicting math competency through neuropsychological interpretation of WISC-III variance components. *Journal of Psychoeducational Assessment, 21,* 358-380. doi: 10.1177/073428290302100404

Hale, J. B., Fiorello, C. A., Kavanagh, J. A., Hoeppner, J. B., & Gaither, R. A. (2001). WISC-III predictors of academic achievement for children with learning disabilities: Are global and factor scores comparable? *School Psychology Quarterly, 16,* 31-55. doi: 10.1521/scpq.16.1.31.19158

Hale, J. B., Fiorello, C. A., Kavanagh, J. A., Holdnack, J. A., & Aloe, A. M. (2007). Is the demise of IQ justified? A response to the special issue authors. *Applied Neuropsychology, 21,* 37-51. doi: 10.1080/09084280701280445

Hanson, J., Sharman, L., & Esparza-Brown, J. (2009). *Patterns of strengths and Weaknesses in specific learning disabilities: What's it all about?* Retrieved from http://www.ospaonline.com/Default.aspx?pageId=417777

Hebb, D. O. (1942). The effects of early and late brain injury upon test scores, and the nature of normal adult intelligence. *Proceedings of the American Philosophical Society, 85,* 275-292. Retrieved from http://www.jstor.org/stable/985007

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA:

Academic Press.

Hempel, C. G. (1965). *Aspects of scientific explanation and other essays in the philosophy of science*. New York: Free Press.

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

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in meta-analysis. *Statistics in Medicine, 21,* 1539-1538. doi: 10.1002/sim.1186

Hopkins, K. D. (1979). Obtaining chi-square tests of association and goodness of fit from proportions and percents. *Journal of Experimental Education, 47,* 380-386. Retrieved from http://www.jstor.org.libproxy.chapman.edu/stable/20151302

Horn, J. L. (1965). *Fluid and crystallized intelligence: A factor analytic study of the structure among primary mental abilities*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign. (AAT 6507113)

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligence. *Journal of Educational Psychology, 57,* 253-270. doi: 10.1037/h0023816

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

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

Jones, W. T. (1952). *A history of western philosophy*. New York: Harcourt-Brace.

Kahana, S. Y., Youngstrom, E. A., & Glutting, J. J. (2002). Factor and subtest discrepancies on the Differential Ability Scales: Examining prevalence and validity in predicting academic achievement. *Assessment, 9,* 82-93. doi: 10.1177/1073191102009001010

Kamphaus, R. W. (2009). Assessment of intelligence and achievement. In T. B. Gutkin & C. R. Reynolds (Eds.), *The handbook of school psychology* (4th ed.; pp. 230-246). Hoboken, NJ: Wiley.

Kamphaus, R. W., Winsor, A. P., Rowe, E. W., & Kim, S. (2005). History of intelligence test interpretation. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues*. (2nd ed., pp. 23-38) New York: Guilford.

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

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

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

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

Keith, T. Z., Fine, J. G., Taub, G. E., Reynolds, M. R., & Kranzler, J. H. (2006). Higher-order, multi-sample, confirmatory factor analysis of the Wechsler Intelligence Scale for Children-Fourth Edition: What does it measure? *School Psychology Review, 35,* 108-127. Retrieved from http://www.nasponline.org

Keith, T. Z., Low, J. A., Reynolds, M. R., Patel, P. G., & Ridley, K. P. (2010). Higher-order factor structure of the Differential Ability Scales-II: Consistency across ages

4 to 17. *Psychology in the Schools, 47,* 676-697. doi: 10.1002/pits.20498

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,* 635-650. doi: 10.1002/pits.20496

Kirk, R. E. (2013). *Experimental design: Procedures for the behavioral science* (4th ed.). Thousand Oaks, CA: Sage.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford.

Kotz, K. M., Watkins, M. W., & McDermott, P. A. (2008). Validity of the general Conceptual Ability score on the Differential Ability Scales as a function of significant and rare interfactor variability. *School Psychology Review, 37,* 261-278. Retrieved from http://www.nasponline.org

Lee, S., & Hershberger, S. (1990). A simple rule for generating equivalent models in covariance structure modeling. *Multivariate Behavioral Research, 25,* 313-334. doi: 10.1207/s15327906mbr2503_4

Light, R., Singer, J., & Willett, J. (1990). *By design: Planning research on higher education*. Cambridge, MA: Harvard University Press.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83,* 1198-1202. Retrieved from http://www.jstor.org/stable/2290157

Livingston, R. B., Jennings, E., Reynolds, C. R., & Gray, R. M. (2003). Multivariate analyses of the profile stability of intelligence tests: High for IQs, low to very low for subtest analyses. *Archives of Clinical Neuropsychology, 18,* 487–507. doi:

10.1016/S0887-6177(02)00147-6

Locke, S., McGrew, K. S., Ford, L. (2011). A multiple group confirmatory factor analysis of the structural invariance of the Cattell-Horn-Carroll theory of cognitive abilities across matched Canadian and U.S. samples. Olympia, WA: Woodcock-Munoz Foundation Press.

Luria, A. R. (1966). *Higher cortical functions in man*. New York: Basic Books.

Mackintosh, N. J. (2011). *IQ and human intelligence* (2nd ed.). New York: Oxford University Press.

Macmann, G. M., & Barnett, D. W. (1997). Myth of the master detective: Reliability of interpretations for Kaufman's "intelligent testing" approach to the WISC-III. *School Psychology Quarterly, 12,* 197-234. doi: 10.1037/h0088959

Major, J. T., Johnson, W., & Bouchard, T. J. (2011). The dependability of the general factor of intelligence: Why small, single-factor models do not adequately represent g. *Intelligence, 39*, 418-433. doi: 10.1016/j.intell.2011.07.002

Matarazzo, J. D., & Herman, D. O. (1985). Clinical uses of the WAIS-R: Base rates of differences between VIQ and PIQ in the WAIS-R standardization sample. In B. B. Wolman (Ed.), *Handbook of Intelligence* (pp. 899-932). New York: John Wiley.

Mather, N., & Wendling, B. J. (2009). Woodcock-Johnson III Tests of Achievement. In J. A. Naglieri & S. Goldstein (Eds.), *Practitioner's guide to assessing intelligence and achievement* (pp. 503-536). Hoboken, NJ: Wiley.

Mather, N., Wendling, B. J., & Woodcock, R. W. (2001). *Essentials of WJ III Tests of Achievement assessment*. New York: Wiley.

Mather, N. & Woodcock, R. W. (2001a). *Woodcock-Johnson III Tests of Achievement examiner's manual*. Itasca, IL: Riverside.

Mather, N., & Woodcock, R. W. (2001b). *Woodcock-Johnson III Tests of Cognitive Abilities examiner's manual*. Itasca, IL: Riverside.

Maynard, J. L., Floyd, R. G., Acklie, T. J., & Houston, L. (2011). General factor loadings and specific effects of the Differential Ability Scales, Second Edition composites. *School Psychology Quarterly, 26,* 108-118. doi: 10.1037/a0023025

McDermott, P. A., Fantuzzo, J. W., & Glutting, J. J. (1990). Just say no to subtest analysis: A critique on Wechsler theory and practice. *Journal of Psychoeducational Assessment, 8,* 290-302. doi: 10.1177/073428299000800307

McDermott, P. A., Fantuzzo, J. W., Glutting, J. J., Watkins, M. W., & Baggaley, A. R. (1992). Illusions of meaning in the ipsative assessment of children's ability. *The Journal of Special Education, 25,* 504-526. doi: 10.1177/002246699202500407

McGill, R. J., & Busse, R. T. (2012, February). *The Incremental Validity of CHC Factors on the KABC-II*. Poster presented at the annual meeting of the National Association of School Psychologists, Seattle, WA.

McGrew, K. (1993). The relationship between the Woodcock-Johnson Psycho-Educational Battery - Revised Gf-Gc cognitive clusters and reading achievement across the life-span. *Journal of Psychoeducational Assessment Monograph Series. WJ-R Monograph,* 39-53.

McGrew, K. S. (2005). The Cattell-Horn-Carroll theory of cognitive abilities: Past, present, and future. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 136-182). New

York: Guilford.

McGrew, K. S. (2008, January, 2). Beyond IQ: A model of academic competence and motivation [Web log post]. Retrieved from http://www.iqscorner.com/2008/01 /beyond-iq-model-of-academic-competence.html

McGrew, K. S. (2009). Editorial: CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37,* 1-10. doi: 10.1016/j.intell.2008.08.004

McGrew, K. S. (2012, September). *Implications of 20 years of CHC cognitive-achievement research: Back to the future and beyond CHC*. Paper presented at the Inaugural session of the Richard Woodcock Institute for Advancement of Contemporary Cognitive Assessment. Medford, MA.

McGrew, K. S., & Evans, J. (2004). *Internal and external factorial extensions to the Cattell-Horn-Carroll (CHC) theory of cognitive abilities: A review of factor analytic research since Carroll's seminal 1993 treatise*. St. Cloud, MN: Institute for Applied Psychometrics.

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

McGrew, K., & Hessler, G. (1995). The relationship between the WJ-R Gf-Gc cognitive clusters and mathematics achievement across the lifespan. *Journal of Psychoeducational Assessment, 13,* 21-38. doi: 10.1177/073428299501300102

McGrew, K., & Knopik, S. (1993). The relationship between the WJ-R Gf-Gc cognitive clusters and writing achievement across the life-span. *School Psychology Review, 22,* 687-695. Retrieved from http://www.nasponline.org

McGrew, K. S., Schrank, F. A., & Woodcock, R. W. (2007). *Woodcock-Johnson III Normative Update technical manual*. Rolling Meadows, IL: Riverside.

McGrew, K. S., & Wendling, B. J. (2010). Cattell-Horn-Carroll cognitive-achievement relations: What we have learned from the last 20 years of research. *Psychology in the Schools, 47,* 651-675. doi: 10.1002/pits.20497

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

Meehl, P. E. (2002). Cliometric metatheory II: Criteria scientists use in theory appraisal and why it is rational to do so. *Psychological Reports, 91,* 339-404. doi: 10.2466/pr0.2002.91.2.339

Morgan, K. E., Rothlisberg, B. A., McIntosh, D. E., & Hunt, M. S. (2009). Confirmatory factor analysis of the KABC-II in preschool children. *Psychology in the Schools, 46,* 515-526. doi: 10.1002/pits.20394

Murray, A. L., Dixon, H. & Johnson, W. (2013). Spearman's law of diminishing returns: A statistical artifact? *Intelligence, 41,* 439-451. doi: 10.1016/j.intell.2013.06.007

Nagle, R. J. (2007). Issues in preschool assessment. In B. A. Bracken & R. J. Nagle (Eds.), *Psychoeducational assessment of the preschool children* (4th ed., pp. 29-48). Mahwah, NJ: Erlbaum.

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

Nelson, J. M., & Canivez, G. L. (2012). Examination of the structural, convergent, and incremental validity of the Reynolds Intellectual Scales (RIAS) with a clinical sample. *Psychological Assessment, 24,* 129-140. doi: 10.1037/a0024878

Nelson, J. M., Canivez, G. L., & Watkins, M. W. (2013). Structural and incremental validity of the Wechsler Adult Intelligence Scale-Fourth Edition with a clinical sample. *Psychological Assessment, 25,* 618-630. doi: 10.1037/a0032086

Newton, J. H., & McGrew, K. S. (2010). Introduction to the special issue: Current research in Cattell-Horn-Carroll-based assessment. *Psychology in the Schools, 47,* 621-634. doi: 10.1002/pits.20495

Noll, J. G., & Horn, J. L. (1998). Age differences in processes of fluid and crystallized intelligence. In J. J. McArdle & R. W. Woodcock (Eds.), *Human cognitive abilities in theory and practice* (pp. 263-282). Mahwah, NJ: Erlbaum.

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

Oh, H., Glutting, J. J., Watkins, M. W., Youngstrom, E. A., & McDermott, P. A. (2004). Correct interpretation of latent versus observed abilities: Implications from structural equation modeling applied to the WISC-III and WIAT linking sample. *Journal of Special Education, 38,* 159-173. doi: 10.1177/00224669040380030301

Parkin, J. R., & Beaujean, A. A. (2012). The effects of Wechsler Intelligence Scale for Children-Fourth Edition cognitive abilities on math achievement. *Journal of School Psychology, 50,* 113-128. doi: 10.1016/j.jsp.2011.08.003

Pedhazur, E. J. (1982). *Multiple regression in behavioral research: Explanation and prediction* (2nd ed.). Orlando, FL: Holt, Rinehart, & Wilson.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). New York: Holt, Rinehart, & Winston.

Pfeifer, S. I., Reddy, L. A., Kletzel. J. E., Schmelzer, E. R., & Boyer, L. M. (2000). The practitioner's view of IQ testing and profile analysis. *School Psychology*

Quarterly, 15, 376-385. doi: 10.1037/h0088795

Pintner, R. (1923). *Intelligence testing*. New York: Holt, Rinehart, & Winston.

Rapaport, D., Gil, M., & Schafer, R. (1945). *Diagnostic psychological testing: The theory, statistical evaluation, and diagnostic application of a battery of tests* (Vol. 1). Chicago: Yearbook Medical.

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment, 12,* 287-297. doi: 10.1037/1040-3590.12.3.287

Reschly, D. J. (2008). School psychology paradigm shift and beyond. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology* (5th ed., pp. 3-15). Bethesda, MD: National Association of School Psychologists.

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

Reynolds, M. R. (2013). Interpreting intelligence test composite scores in light of Spearman's law of diminishing returns. *School Psychology Quarterly. 28,* 63-76. doi: 10.1037/spq0000013

Reynolds, M. R., Hajovsky, D. B., Niileksela, C. R., & Keith, T. Z. (2011). Spearman's Law of Diminishing Returns and the DAS-II: Do g effects on subtest scores depend on g? *School Psychology Quarterly, 26,* 275–289. doi: 10.1037/a0026190

Reynolds, M. R., & Ketih, T. Z. (2013). Measurement and statistical issues in child assessment research. In D. H. Saklofske, C. R. Reynolds, & V. L. Schwean (Eds.), *The oxford handbook of child psychological assessment* (pp. 48-83). New York: Oxford University Press.

Reynolds, M. R., Keith, T. Z., & Beretvas, S. N. (2010). Use of factor mixture modeling to capture Spearman’s law of diminishing returns. *Intelligence, 38,* 231–241. doi: 10.1016/j.intell.2010.01.002

Roid, G. H. (2003). *Stanford-Binet Intelligence Scales-Fifth Edition*. Austin, TX: Pro-Ed. Salvia, J., & Ysseldyke, J. E. (2007). *Assessment in special and inclusive education* (10th ed.). Boston: Houghton Mifflin

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63,* 581-592. doi: 10.1093/biomet/63.3.581

Sanders, S., McIntosh, D. E., Dunham, M., Rothlisberg, B. A., & Finch, H. (2007). Joint confirmatory factor analysis of the Differential Ability Scales and the Woodcock-Johnson Tests of Cognitive Abilities-Third Edition. *Psychology in the Schools, 44,* 119-138. doi: 10.1002/pits.20211

Sandoval, J. (2003). Review of the Woodcock-Johnson III. In B. S. Plake & J. C. Impara (Eds.), *The fifteenth mental measurements yearbook* (pp. 1024-1028). Lincoln, NE: Buros Institute of Mental Measurements.

Sapp, M. (2004). Confidence intervals within hypnosis research. *Sleep and Hypnosis, 6,* 169-176. Retrieved from http://search.proquest.com/docview/197749389?accountid=10051

Sattler, J. M. (2008). *Assessment of children: Cognitive foundations* (5th ed.). San Diego, CA: Jerome M. Sattler.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147-177. doi:10.1037//1082-989X.7.2.147

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika, 22,* 53-61. doi: 10.1007/BF02289209

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

Schneider, J., & McGrew, K. (2011). *AP #10: "Just say no" to averaging IQ subtest scores*. St Cloud, MN: Institute for Applied Psychometrics.

Schneider, W. J. (2008). Playing statistical Ouija board with communality analysis: Good questions, wrong assumptions. *Applied Neuropsychology, 15,* 44-53. doi: 10.1080/09084280801917566

Schneider, W. J. (2011, June 30). General intelligence: To g or not to g? [Web log post]. Retrieved from http://www.iqscorner.com/2011/06/general-intelligence-to-g-or-not-to-g.html

Schneider, W. J. (2013). What if we took our models seriously? Estimating latent scores in individuals. *Journal of Psychoeducational Assesment, 31,* 186-201. doi: 10.1177/0734282913478046

Schneider, W. J., & 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: Guilford.

Schrank, F. A., Flanagan, D. P., Woodcock, R. W., & Mascolo, J. T. (2002). *Essentials of WJ III cognitive abilities assessment*. New York: John Wiley.

Schrank, F. A., Miller, D. C., Wendling, B. J., & Woodcock, R. W. (2010). *Essentials of WJ III cognitive abilities assessment* (2nd ed.). Hoboken, NJ: John Wiley.

Sechrest, L. (1963). Incremental validity: A recommendation. *Educational and Psychological Measurement, 23,* 153-158. doi: 10.1177/001316446302300113

Shapiro, E. S. (2011). *Academic skills problems: Direct assessment and intervention* (4th ed.). New York: Guilford.

Snow, R. E., Kyllonen, P. C., & Marshalek, B. (1984). The topography of ability and learning correlations. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 2., pp. 47-103). Hillsdale, NJ: Erlbaum.

Spearman, C. (1904). “General intelligence,” objectively determined and measured. *American Journal of Psychology, 15,* 201-293. Retrieved from http://psychclassics.yorku.ca/Spearman

Spearman, C. (1927). *The abilities of man: Their nature and measurement*. New York: Macmillan.

Spearman, C., & Jones, L. W. (1950). *Human ability: A continuation of “The Abilities of Man.”* London: Macmillan.

Special Education Local Plan Area Administrators of California. (2010). *New approaches to comprehensive assessment for SLD eligibility part two: Patterns of strengths and weaknesses*. Retrieved from http://www.bcoe.org/SELPA/Resources-Info/SLD%20Eligibility/SLD_Eligibility-PSW_3-29-10.pdf

Spencer, H. (1855). *The principles of psychology*. London: Longman, Brown, Green, & Longmans.

Sternberg, R. J. (1984). A contextualist view of the nature of intelligence. *International Journal of Psychology*, *19,* 307-334. doi: 10.1080/00207598408247535

Stevens, J. P. (2009). *Applied multivariate statistics for the social sciences* (5th ed.). New York: Routledge.

Suen, H. K., & French, J. L. (2003). A history of the development of psychological and educational testing. In C. R. Reynolds & R.W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Intelligence, aptitude, and achievement* (2nd ed., pp. 3-23). New York: Guilford Press.

Swanson, H. L.,, Zheng, X., & Jerman, O. (2009). Working memory, short-term memory, and reading disabilities: A selective meta-analysis of the literature. *Journal of Learning Disabilities, 42,* 260-287. doi: 10.1177/0022219409331958

Tabachnick, B. G., & Fiddell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Allyn & Bacon.

Taub, G. E., Keith, T. Z., Floyd, R. G., & McGrew, K. S. (2008). Effects of general and broad cognitive abilities on mathematics achievement. *School Psychology Quarterly, 23,* 187-198. doi: 10.1037/1045-3830.23.2.187

Taub, G. E., & McGrew, K. S. (2004). A confirmatory factor analysis of Cattell-Horn-Carroll theory and cross-age invariance on the Woodcock-Johnson Tests of Cognitive Abilities III. School Psychology Quarterly, 19, 72-87. doi: 10.1521/scpq.19.1.72.29409

Terman, L. M. (1921). Intelligence and its measurement: A symposium. *Journal of Educational Psychology, 12,* 127-133. doi: 10.1037/h0076078

Thompson, B. (2000). Ten commandments of structural equation modeling. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 261-284). Washington, DC: American Psychological Association.

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools, 44,* 423-432. doi: 10.1002/pits.20234

Thomson, G. H. (1916). A hierarchy without a general factor. *British Journal of*

Psychology, 8, 271-281. doi: 10.1111/j.2044-8295.1916.tb00133.x

Thorndike, R. L., Lay, W., & Dean, P. R. (1909). The relation of accuracy in sensory discrimination to general intelligence. *American Journal of Psychology, 20,* 364-369. Retrieved from http://www.jstor.org/stable/1413366

Thorndike, R. M., & Lohman, D. F. (1990). *A century of ability testing*. Chicago: Riverside.

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

Thurstone, L. L. (1947). *Multiple factor analysis: A development and expansion of the vectors of the mind*. Chicago: University of Chicago Press.

Traub, R. E. (1991). *Reliability for the social sciences*. Newbury Park, CA: Sage.

Tucker-Drob, E. M. (2009). Differentiation of cognitive abilities across the lifespan. *Developmental Psychology, 45,* 1097-1118. doi: 10.1037/a0015864

Valencia, R. R., & Suzuki, L. A. (2001). *Intelligence assessment and minority students: Foundations, performance factors, and assessment issues*. Thousand Oaks, CA: Sage.

Vanderwood, M. L., McGrew, K. S., Flanagan, D. P., & Keith, T. Z. (2001). The contribution of general and specific cognitive abilities to reading achievement. *Learning and Individual Differences, 13,* 159-188. doi: 10.1016/S1041-6080(02)00077-8

Verhaeghen, P., & Salthouse, T. A. (1997). Meta-analyses of age-cognition relations in adulthood: Estimates of linear and non-linear age effects and structural models. *Psychological Bulletin, 122,* 231-249. doi: 10.1037/033-2909.122.3.231

Vernon, P. E. (1950). *The structure of human abilities*. London: Methuen.

Wasserman, J. D. (2012). A history of intelligence assessment: The unfinished tapestry. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 3-54). New York: Guilford Press.

Watkins, M. W. (2000). Cognitive profile analysis: A shared professional myth. *School Psychology Quarterly, 15,* 465-479. doi: 10.1037/h0088802

Watkins, M. W. (2006). Orthogonal higher-order structure of the WISC-IV. *Psychological Assessment, 16,* 123-125. doi: 10.1037/1040-3590.18.1.123

Watkins, M. W., & Canivez, G. L. (2004). Temporal stability of WISC-III subtest composite strengths and weaknesses. *Psychological Assessment, 16,* 133–138. doi: 10.1037/1040-3590.16.2.133

Watkins, M. W., & Glutting, J. J. (2000). Incremental validity of the WISC-III profile elevation, scatter, and shape information for predicting reading and math achievement. *Psychological Assessment, 12,* 402-408. doi: 10.1037/1040-3590.12.4.402

Watkins, M. W., Glutting, J. J., & Youngstrom, E. A. (2005). Issues in subtest profile analysis. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 251-268). New York: Guilford Press.

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

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

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

Weiss, L. G., Saklofske, D. M., Schwartz, D. M., Prifitera, A, & Courville, T. (2006). Advanced clinical interpretation of WISC-IV index scores. In L. G. Weiss, D. H. Saklofske, A. Prifitera, & J. A. Holdnack (Eds.), *WISC-IV advanced clinical interpretation* (pp. 140-181). San Diego, CA: Academic Press.

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

Woodcock, R. W. (1990). Theoretical foundations of the WJ-R measures of cognitive ability. *Journal of Psychoeducational Assessment, 8,* 231-258. doi:10.1177/073428299000800303

Woodcock, R. W., & Dahl, M. N. (1971). *A common scale for the measurement of person ability and test item difficulty* (AGS Paper No. 10). Circle Pines, MN: American Guidance Service.

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson Psycho-Educational Battery-Revised*. Chicago: Riverside.

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

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

Woodcock, R. W., McGrew, K. S., Mather, N. (2001c). *Woodcock-Johnson III Tests of Cognitive Abilities*. Itasca, IL: Riverside.

Woodcock, R. W., McGrew, K. S., & Schrank, F. A., Mather, N. (2007). *Woodcock-Johnson III Normative Update*. Rolling Meadows, IL: Riverside

Youngstom, E. A., Kogos, J. L., & Glutting, J. J. (1999). Incremental efficacy of

Differential Ability Scales factor scores in predicting individual achievement criteria. *School Psychology Quarterly, 14,* 26-39. doi: 10.1037/h0088996

Table 1

Demographic Information for the Study Sample (N = 4,722)

Variable	*n*	Percent of Sample	Percent of U.S. Population
Sex			
Male	2382	50.4	51.2
Female	2340	49.6	48.8
Race			
Caucasian American	3702	78.4	78.5
African American	684	14.5	16.1
American Indian	95	2.0	1.3
Asian American	241	5.1	4.1
Hispanic/Latino/Chicano			
Yes	567	12.0	18.7
No	4155	88.0	81.3
Census Region			
Northeast	1133	24.0	17.8
Midwest	978	20.7	22.3
South	1487	31.5	35.9
West	1124	23.8	24.0
Community Size			
Large City	2800	59.3	68.3
Suburban	1025	21.7	10.7
Rural	897	19.0	21.0
Type of School			
Public	4099	86.8	86.5
Private	571	12.1	11.3
Home	52	1.1	2.2
Foreign Born			
No	4486	95.0	94.3
Yes	236	5.0	5.7
Father's Education			
Less than High School	527	11.7	13.3
High School	1507	33.5	31.8
More than High School	2465	54.8	54.9
Not Available	223		
Mother's Education			
Less than High School	432	9.6	10.9
High School	1485	33.0	29.5
More than High School	2583	57.4	59.6
Not Available	222		

Table 2

Frequency Distribution of Cases across Grade and Age

Grade	*n*	Age	*n*
K	121	6	308
1	325	7	335
2	356	8	431
3	490	9	533
4	574	10	579
5	552	11	428
6	368	12	352
7	338	13	324
8	328	14	292
9	285	15	302
10	291	16	308
11	276	17	249
12	232	18	281
12+	132		
Not Available	54[1]		

Note. [1]Cases were not included in the analysis for research question 2.

Table 3

WJ-COG CHC Factor g Loadings for Ages 6 to 18

Age	Gc	Gf	Ga	Gv	Glr	Gsm	Gs
6	.22	.18	.11	.08	.16	.13	.11
7	.22	.18	.10	.08	.16	.13	.12
8	.22	.19	.10	.08	.16	.13	.12
9	.22	.19	.10	.08	.16	.13	.12
10	.24	.20	.09	.08	.15	.13	.12
11	.24	.20	.09	.08	.15	.13	.12
12	.24	.20	.10	.08	.15	.13	.12
13	.24	.20	.10	.08	.15	.13	.12
14.	.24	.20	.10	.08	.15	.13	.12
15	.24	.20	.10	.08	.15	.13	.11
16	.23	.20	.10	.08	.14	.13	.11
17	.23	.20	.10	.08	.14	.13	.11
18	.23	.20	.10	.08	.13	.13	.11
Mean	.23	.19	.10	.08	.15	.13	.12

Note. Estimated factor weights are a combination of individual subtest loadings on the first principal component as reported in the WJ-III NU technical manual (McGrew, Schrank, & Woodcock, 2007).

Table 4

Mean Internal Consistency Estimates for the WJ-III Variables (Ages 6 to 18)

Variables	Reliability Coefficient
Tests of Cognitive Abilities	
GIA	.97
Crystallized Ability	.94
Fluid Reasoning	.95
Auditory Processing	.88
Visual Processing	.78
Long-Term Retrieval	.87
Short-Term Memory	.86
Processing Speed	.91
Tests of Achievement	
Broad Reading	.96
Broad Mathematics	.95
Broad Written Language	.92
Basic Reading Skills	.93
Reading Comprehension	.91
Math Calculation	.90
Math Reasoning	.94
Written Expression	.84
Oral Expression	.82
Listening Comprehension	.85

Note. The WJ-III NU Technical Manual (McGrew, Schrank, & Woodcock, 2007) reports that Rasch procedures were utilized to calculate reliability coefficients for tests that involve speed and employ multiple point formats; all remaining tests utilized the split-half method.

Table 5

WJ-COG Independent Variable Descriptions (N = 8)

Variable Name	Type	Description
General Intellectual Ability (GIA)	Standard Score Aggregate	Representation of the first principle component, or single *g* factor accounting for the most variance in overall performance on the tests that comprise the scale
Comprehension-Knowledge (Gc)	Standard Score Index	Includes the breadth and depth of a person's acquired knowledge, the ability to communicate one's knowledge, and the ability to reason using previously learned experiences and procedures
Long-Term Retrieval (Glr)	Standard Score Index	The ability to store information and fluently retrieve it later in the process of thinking
Visual-Spatial Thinking (Gv)	Standard Score Index	The ability to perceive, analyze, synthesize, and think with visual patterns, including the ability to store and recall visual

(Continued)

		representations
Auditory Processing (Ga)	Standard Score Index	The ability to analyze, synthesize, and discriminate auditory stimuli, including the ability to process and discriminate speech sounds that may be presented under distorted conditions
Fluid Reasoning (Gf)	Standard Score Index	The broad ability to reason, form concepts, and solve problems using unfamiliar information or novel procedures
Processing Speed (Gs)	Standard Score Index	The ability to perform automatic cognitive tasks, particularly when measured under pressure to maintain focused attention
Short-Term Memory (Gsm)	Standard Score Index	The ability to apprehend and hold information in immediate awareness and then use it within a few seconds

Note. All independent variables from the WJ-COG (Woodcock, McGrew, & Mather, 2001). All variable descriptions from (Mather & Woodcock, 2001b).

Table 6

WJ-ACH Dependent Variable Descriptions (N = 10)

Variable Name	Type	Description
Broad Reading	Standard Score Composite	Provides a broad measure of reading achievement
Broad Mathematics	Standard Score Composite	Provides a broad measure of math achievement
Broad Written Language	Standard Score Composite	Provides a broad measure of written language
Listening Comprehension	Standard Score Cluster	Aggregate measure of listening ability and verbal comprehension
Oral Expression	Standard Score Cluster	Aggregate measure of linguistic competency and expressive vocabulary
Basic Reading Skills	Standard Score Cluster	Aggregate measure of sight vocabulary, phonics, and structural analysis
Reading Comprehension	Standard Score Cluster	Aggregate measure of comprehension, vocabulary, and reasoning

(Continued)

Math Calculation Skills	Standard Score Cluster	Aggregate measure of computational skills and automaticity with basic math facts and provides a measure of basic mathematics skills
Math Reasoning	Standard Score Cluster	Aggregate measure of problem solving, analysis, reasoning, and vocabulary
Written Expression	Standard Score Cluster	Aggregate measure of meaningful written expression and fluency providing a measure of written expression skills

Note. All independent variables from the WJ-ACH (Woodcock, McGrew, & Mather, 2001). All variable descriptions from (Mather & Woodcock, 2001a).

Table 7

Univariate Descriptive Statistics for Study Variables

Variables	*N*	*M*	*SD*	Skewness	Kurtosis
GIA	2130	100.36	14.94	-.12	.39
Crystallized Ability	2903	100.85	14.86	-.40	.53
Fluid Reasoning	3253	100.30	15.45	-.34	.32
Auditory Processing	3435	100.25	15.64	.11	.40
Visual Processing	2775	100.36	14.62	-.15	.37
Long-Term Retrieval	3078	100.16	14.88	.03	.34
Short-Term Memory	3746	100.74	15.49	-.07	.48
Processing Speed	3326	99.99	15.11	-.07	.47
Basic Reading Skills	4028	100.94	15.00	-.35	.57
Reading Comprehension	3217	101.15	15.53	-.37	1.12
Math Calculation Skills	3961	100.19	15.63	-.25	.70
Math Reasoning	3647	100.31	15.82	-.11	.40
Written Expression	3890	101.13	15.10	-.32	.90
Oral Expression	3183	100.06	15.30	-.11	.43
Listening Comprehension	3831	100.00	16.08	-.27	.46
Broad Reading	3845	101.38	15.26	-.36	.99
Broad Mathematics	3954	100.76	15.70	-.20	.68
Broad Written Language	3877	100.51	15.44	-.41	1.07

Note. Obtained values rounded to the nearest hundredth.

Table 8

Tests of Normality

Variable	Kolmogorov-Smirnov	Shapiro-Wilk
GIA	.020	.996*
Crystallized Ability	.033*	.990*
Fluid Reasoning	.034*	.990
Auditory Processing	.028	.995*
Visual Processing	.022	.998
Long-Term Retrieval	.022	.997
Short-Term Memory	.017	.998
Processing Speed	.022	.997
Basic Reading Skills	.029	.988*
Reading Comprehension	.044*	.982*
Math Calculation Skills	.021	.993*
Math Reasoning	.025	.996*
Written Expression	.033*	.990*
Oral Expression	.022	.996
Listening Comprehension	.028	.996*
Broad Reading	.039*	.985
Broad Mathematics	.027	.991*
Broad Written Language	.032*	.987*

Note. * $p < .001$.

Table 9

Bivariate Correlational Coefficients between Independent Variables

Variable	GIA	Gc	Gf	Ga	Gv	Glr	Gsm	Gs
GIA	-							
Crystallized Ability (Gc)	.84	-						
Fluid Reasoning (Gf)	.82	.60	-					
Auditory Processing (Ga)	.63	.50	.41	-				
Visual Processing (Gv)	.49	.35	.37	.25	-			
Long-Term Retrieval (Glr)	.71	.57	.56	.42	.37	-		
Short-Term Memory (Gsm)	.72	.45	.47	.41	.26	.45	-	
Processing Speed (Gs)	.53	.31	.36	.33	.28	.40	.33	-

Note. Values rounded to nearest hundredth. All coefficients were statistically significant (p < .01, two-tailed).

Table 10

Collinearity Diagnostics

Variable	Tolerance	VIF	Condition
	Broad Reading		
GIA	**.015**	**67.12**	21.09
Crystallized Ability	**.097**	**10.27**	23.47
Fluid Reasoning	.115	8.72	24.64
Auditory Processing	.418	2.39	25.48
Visual Processing	.618	1.62	29.36
Long-Term Retrieval	.414	2.41	**32.11**
Short-Term Memory	.156	6.41	**32.90**
Processing Speed	.485	2.06	**199.65***
	Broad Mathematics		
GIA	**.015**	**68.37**	21.06
Crystallized Ability	**.097**	**10.33**	23.45
Fluid Reasoning	.114	8.79	24.61
Auditory Processing	.412	2.43	25.35
Visual Processing	.620	1.61	29.38
Long-Term Retrieval	.411	2.44	**31.92**
Short-Term Memory	.153	6.55	**32.88**
Processing Speed	.479	2.09	**199.72***
	Broad Written Language		
GIA	**.015**	**67.67**	21.04
Crystallized Ability	**.098**	**10.23**	23.44
Fluid Reasoning	.114	8.74	24.53
Auditory Processing	.417	2.40	25.30
Visual Processing	.621	1.61	29.24
Long-Term Retrieval	.411	2.43	**31.89**
Short-Term Memory	.153	6.53	**32.84**
Processing Speed	.480	2.08	**199.00***

Note. VIF = variance inflation factor. **Bold** indicates significant value. * Indicates high variance proportion with two or more variables.

Table 11

Missing Data Analysis (N = 4,722)

Variable	Valid *N*	Missing	Percent Missing
GIA	2130	2592	55
Crystallized Ability	2903	1819	39
Fluid Reasoning	3253	1469	32
Auditory Processing	3435	1287	27
Visual Processing	2775	1947	41
Long-Term Retrieval	3078	1644	35
Short-Term Memory	3746	976	21
Processing Speed	3326	1396	30
Basic Reading Skills	4028	694	15
Reading Comprehension	3217	1505	32
Math Calculation Skills	3961	761	16
Math Reasoning	3647	1075	23
Written Expression	3890	832	18
Oral Expression	3183	1539	33
Listening Comprehension	3831	891	19
Broad Reading	3845	877	19
Broad Mathematics	3954	768	16
Broad Written Language	3877	845	18

Note. Percentages rounded to nearest tenth. According to Enders (2002), missing data percentages for individual variables which exceed 5% are considered problematic.

Table 12

Multiple Imputation Parameter Estimates

Variable	*N*	*M*	*SD*	Hedge's *g*[1]
GIA	25,575	100.23	14.15	-.009
Crystallized Ability	26,348	100.73	14.28	-.008
Fluid Reasoning	26,698	100.45	15.22	.009
Auditory Processing	26,880	100.29	15.45	.003
Visual Processing	26,220	100.22	14.47	-.009
Long-Term Retrieval	26,523	100.41	14.50	.017
Short-Term Memory	27,191	100.83	15.26	.011
Processing Speed	26,771	100.16	14.92	.003
Basic Reading Skills	27,473	100.81	14.87	-.008
Reading Comprehension	26,662	100.71	15.39	-.028
Math Calculation Skills	27,406	100.10	15.27	-.005
Math Reasoning	27,092	100.36	15.52	.003
Written Expression	27,335	101.07	14.87	-.004
Oral Expression	26,628	100.36	14.95	.020
Listening Comprehension	27,276	99.62	15.78	-.003
Broad Reading	27,290	100.92	15.15	-.030
Broad Mathematics	27,399	100.63	15.33	-.008
Broad Written Language	27,322	100.40	15.22	-.007

Note. [1] Standardized means difference test against obtained parameter estimates from original sample. Valid listwise (*N* = 25,147).

Table 13

Hierarchical Multiple Regression Analysis Predicting Basic Reading Skills (n = 2,129[a]) from the WJ-COG General Intellectual Ability Composite and CHC Broad Factor Variables

Predictor	*F*	F_{inc}	R^2	95% CI	Δ R^2	Incremental Variance[c]
GIA	1802.68*	-	.46	.43-.49	-	46%
CHC Factor Scores (*df* = 7)[b]	249.67*	15.51*	.49	.45-.52	.03	3%
Crystallized Ability	933.72*	35.51*	.47	.44-.50	.01	1%
Fluid Reasoning	960.71*	64.73*	.48	.44-.51	.02	2%
Auditory Processing	904.90*	4.31*	.46	.43-.49	.00	0%
Visual Processing	912.34*	12.37*	.46	.43-.49	.00	0%
Long-Term Retrieval	901.01*	.10	.46	.43-.49	.00	0%
Short-Term Memory	910.45*	10.32*	.46	.43-.49	.00	0%
Processing Speed	901.18*	.28	.46	.43-.49	.00	0%

Note. [a] Subsample utilizing listwise deletion. [b] Degrees of freedom reflects controlling for GIA. Squared multiple correlation coefficients for the CHC factors reflect the total variance accounted for by that variable when it is included in a prediction equation in addition to the GIA. [c] Represents proportion of variance accounted for by variables at their point of entry into the regression equation after controlling for the effects of the GIA. All values rounded to the nearest hundredth, may not sum to 100 due to rounding errors. * $p < .05$.

Table 14

Hierarchical Multiple Regression Analysis Predicting Reading Comprehension (n = 1,813 [a]) from the WJ-COG General Intellectual Ability Composite and CHC Broad Factor Variables

Predictor	*F*	*F*inc	*R*²	95% CI	Δ *R*²	Incremental Variance[c]
GIA	2155.69*	-	.54	.51-.57	-	54%
CHC Factor Scores (*df* = 7)[b]	337.19*	38.87*	.60	.57-.63	.06	6%
Crystallized Ability	1308.67*	231.91*	.59	.56-.62	.05	5%
Fluid Reasoning	1080.11*	21.08*	.54	.51-.57	.01	1%
Auditory Processing	1057.69*	.39	.54	.51-.57	.00	0%
Visual Processing	1072.50*	14.06*	.54	.51-.57	.00	0%
Long-Term Retrieval	1071.09*	12.76*	.54	.51-.57	.00	0%
Short-Term Memory	1090.02*	30.22*	.55	.52-.58	.01	1%
Processing Speed	1063.46*	5.72*	.54	.51-.57	.00	0%

Note. [a] Subsample utilizing listwise deletion. [b] Degrees of freedom reflects controlling for GIA. Squared multiple correlation coefficients for the CHC factors reflect the total variance accounted for by that variable when it is included in a prediction equation in addition to the GIA. [c] Represents proportion of variance accounted for by variables at their point of entry into the regression equation after controlling for the effects of the GIA. All values rounded to the nearest hundredth, may not sum to 100 due to rounding errors. * $p < .05$.

Table 15

Hierarchical Multiple Regression Analysis Predicting Math Calculation Skills (n = 2,106[a]) from the WJ-COG General Intellectual Ability Composite and CHC Broad Factor Variables

Predictor	*F*	F_{inc}	*R*²	95% CI	Δ *R*²	Incremental Variance[c]
GIA	842.03*	-	.29	.25-.32	-	29%
CHC Factor Scores (*df* = 7)[b]	138.47*	27.40*	.35	.31-.38	.06	6%
Crystallized Ability	425.25*	6.34*	.29	.26-.32	.00	0%
Fluid Reasoning	422.56*	2.49	.29	.25-.32	.00	0%
Auditory Processing	425.25*	6.34*	.29	.26-.32	.00	0%
Visual Processing	422.17*	1.94	.29	.25-.32	.00	0%
Long-Term Retrieval	421.91*	1.57	.29	.25-.32	.00	0%
Short-Term Memory	421.22*	.58	.29	.25-.32	.00	0%
Processing Speed	516.45*	136.61*	.33	.30-.36	.04	4%

Note. [a] Subsample utilizing listwise deletion. [b] Degrees of freedom reflects controlling for GIA. Squared multiple correlation coefficients for the CHC factors reflect the total variance accounted for by that variable when it is included in a prediction equation in addition to the GIA. [c] Represents proportion of variance accounted for by variables at their point of entry into the regression equation after controlling for the effects of the GIA. All values rounded to the nearest hundredth, may not sum to 100 due to rounding errors. * $p < .05$.

Table 16

Hierarchical Multiple Regression Analysis Predicting Math Reasoning (n = 2,127[a]) from the WJ-COG General Intellectual Ability Composite and CHC Broad Factor Variables

Predictor	*F*	F_{inc}	R^2	95% CI	Δ R^2	Incremental Variance[c]
GIA	2482.89*	-	.54	.51-.57	-	54%
CHC Factor Scores (*df* = 7)[b]	339.21*	15.74*	.56	.53-.59	.02	2%
Crystallized Ability	1269.03*	25.98*	.54	.52-.57	.01	1%
Fluid Reasoning	1266.70*	23.83*	.54	.52-.57	.01	1%
Auditory Processing	1274.38*	30.91*	.55	.52-.57	.01	1%
Visual Processing	1241.41*	.50	.54	.51-.57	.00	0%
Long-Term Retrieval	1241.44*	.54	.54	.51-.57	.00	0%
Short-Term Memory	1252.71*	10.93*	.54	.51-.57	.00	0%
Processing Speed	1240.87*	.00	.54	.51-.57	.00	0%

Note. [a] Subsample utilizing listwise deletion. [b] Degrees of freedom reflects controlling for GIA. Squared multiple correlation coefficients for the CHC factors reflect the total variance accounted for by that variable when it is included in a prediction equation in addition to the GIA. [c] Represents proportion of variance accounted for by variables at their point of entry into the regression equation after controlling for the effects of the GIA. All values rounded to the nearest hundredth, may not sum to 100 due to rounding errors. * $p < .05$.

Table 17

Hierarchical Multiple Regression Analysis Predicting Written Expression (n = 2,063 [a]) from the WJ-COG General Intellectual Ability Composite and CHC Broad Factor Variables

Predictor	*F*	F_{inc}	R^2	95% CI	Δ R^2	Incremental Variance[c]
GIA	1430.80*	-	.41	.38-.44	-	41%
CHC Factor Scores (*df* = 7)[b]	224.44*	31.16*	.47	.43-.50	.06	6%
Crystallized Ability	727.71*	14.94*	.41	.38-.45	.00	0%
Fluid Reasoning	732.00*	20.00*	.42	.38-.45	.01	1%
Auditory Processing	722.30*	8.55*	.41	.38-.44	.00	0%
Visual Processing	718.06*	3.54	.41	.38-.44	.00	0%
Long-Term Retrieval	718.10*	3.59	.41	.38-.44	.00	0%
Short-Term Memory	715.47*	.50	.41	.38-.44	.00	0%
Processing Speed	782.80*	79.98*	.43	.40-.46	.02	2%

Note. [a] Subsample utilizing listwise deletion. [b] Degrees of freedom reflects controlling for GIA. Squared multiple correlation coefficients for the CHC factors reflect the total variance accounted for by that variable when it is included in a prediction equation in addition to the GIA. [c] Represents proportion of variance accounted for by variables at their point of entry into the regression equation after controlling for the effects of the GIA. All values rounded to the nearest hundredth, may not sum to 100 due to rounding errors. * $p < .05$.

Table 18

Hierarchical Multiple Regression Analysis Predicting Oral Expression (n = 2,126[a]) from the WJ-COG General Intellectual Ability Composite and CHC Broad Factor Variables

Predictor	*F*	F_{inc}	*R*²	95% CI	Δ *R*²	Incremental Variance[c]
GIA	1906.55*	-	.47	.44-.50	-	47%
CHC Factor Scores (*df* = 7)[b]	639.93*	242.35*	.71	.69-.73	.23	23%
Crystallized Ability	2479.39*	1608.92*	.70	.68-.72	.23	23%
Fluid Reasoning	1073.54*	127.23*	.50	.47-.53	.03	3%
Auditory Processing	953.10*	.29	.47	.44-.50	.00	0%
Visual Processing	973.05*	21.32*	.48	.45-.51	.01	1%
Long-Term Retrieval	952.85*	.03	.47	.44-.50	.00	0%
Short-Term Memory	1108.74*	164.33*	.51	.48-.54	.04	4%
Processing Speed	1051.55*	104.06*	.50	.47-.53	.03	3%

Note. [a] Subsample utilizing listwise deletion. [b] Degrees of freedom reflects controlling for GIA. Squared multiple correlation coefficients for the CHC factors reflect the total variance accounted for by that variable when it is included in a prediction equation in addition to the GIA. [c] Represents proportion of variance accounted for by variables at their point of entry into the regression equation after controlling for the effects of the GIA. All values rounded to the nearest hundredth, may not sum to 100 due to rounding errors. * $p < .05$.

Table 19

Hierarchical Multiple Regression Analysis Predicting Listening Comprehension (n = 2,130[a]) from the WJ-COG General Intellectual Ability Composite and CHC Broad Factor Variables

Predictor	F	F_{inc}	R^2	95% CI	ΔR^2	Incremental Variance[c]
GIA	2753.28*	-	.56	.54-.59	-	56%
CHC Factor Scores (*df* = 7)[b]	419.11*	37.91*	.61	.59-.64	.05	5%
Crystallized Ability	1497.41*	105.86*	.59	.56-.61	.02	2%
Fluid Reasoning	1402.84*	23.41*	.57	.54-.60	.01	1%
Auditory Processing	1377.07*	.94	.56	.54-.59	.00	0%
Visual Processing	1388.74*	11.11*	.57	.54-.59	.00	0%
Long-Term Retrieval	1377.12*	.98	.56	.54-.59	.00	0%
Short-Term Memory	1436.68*	52.91*	.58	.55-.60	.01	1%
Processing Speed	1376.51*	.45	.56	.54-.59	.00	0%

Note. [a] Subsample utilizing listwise deletion. [b] Degrees of freedom reflects controlling for GIA. Squared multiple correlation coefficients for the CHC factors reflect the total variance accounted for by that variable when it is included in a prediction equation in addition to the GIA. [c] Represents proportion of variance accounted for by variables at their point of entry into the regression equation after controlling for the effects of the GIA. All values rounded to the nearest hundredth, may not sum to 100 due to rounding errors. * $p < .05$.

Table 20

Hierarchical Multiple Regression Analyses Predicting Broad Achievement Outcomes from the WJ-COG General Intellectual Ability Composite and CHC Broad Factor Variables across Levels of Schooling

K-5	Broad Reading (*n* = 845)[a]				Broad Mathematics (*n* = 896)[a]				Broad Written Language (*n* = 859)[a]			
Predictor	R^2	95% CI	ΔR^2	Inc.[c]	R^2	95% CI	ΔR^2	Inc.[c]	R^2	95% CI	ΔR^2	Inc.[c]
GIA	.48*	.44-.53	-	48%	.45*	.40-.50	-	45%	.41*	.36-.46	-	41%
CHC Factors (*df* = 7)[b]	.53*	.49-.58	.05	5%	.49*	.44-.53	.03	3%	.46*	.41-.51	.05	5%
Gc, Gf, Ga, Gv, Glr												
Gsm, Gs												

6-8	Broad Reading (*n* = 524)[a]				Broad Mathematics (*n* = 521)[a]				Broad Written Language (*n* = 516)[a]			
Predictor	R^2	95% CI	ΔR^2	Inc.[c]	R^2	95% CI	ΔR^2	Inc.[c]	R^2	95% CI	ΔR^2	Inc.[c]
GIA	.55*	.50-.61	-	55%	.49*	.43-.55	-	49%	.46*	.40-.52	-	46%
CHC Factors (*df* = 7)[b]	.62*	.57-.67	.07	7%	.52*	.47-.58	.03	3%	.52*	.46-.57	.06	6%
Crystallized Ability												
Gc, Gf, Ga, Gv, Glr												
Gsm, Gs												

9-12	Broad Reading (*n* = 616)[a]				Broad Mathematics (*n* = 615)[a]				Broad Written Language (*n* = 613)[a]			
Predictor	R^2	95% CI	ΔR^2	Inc.[c]	R^2	95% CI	ΔR^2	Inc.[c]	R^2	95% CI	ΔR^2	Inc.[c]
GIA	.64*	.60-.69	-	64%	.47*	.41-.53	-	47%	.55*	.50-.60	-	55%
CHC Factors (*df* = 7)[b]	.70*	.67-.74	.06	6%	.51*	.45-.56	.04	4%	.59*	.54-.64	.03	3%
Gc, Gf, Ga, Gv, Glr												
Gsm, Gs												

Total Sample	Broad Reading (*n* = 2,059)[a]				Broad Mathematics (*n* = 2,106)[a]				Broad Written Language (*n* = 2,062)[a]			
Predictor	R^2	95% CI	ΔR^2	Inc.[c]	R^2	95% CI	ΔR^2	Inc.[c]	R^2	95% CI	ΔR^2	Inc.[c]
GIA	.55*	.52-.58	-	55%	.46*	.43-.49	-	46%	.46*	.42-.49	-	46%
CHC Factors (*df* = 7)[b]	.60*	.58.-63	.06	6%	.49*	.46-.52	.03	3%	.51*	.48-.54	.05	5%
Gc, Gf, Ga, Gv, Glr												
Gsm, Gs												

Note. Gc = Crystallized Ability, Gf = Fluid Reasoning, Ga = Auditory Processing, Gv = Visual Processing, Glr = Long-Term Retrieval, Gsm = Short-Term Memory, Gs = Processing Speed. [a] Subsample utilizing listwise deletion. Squared multiple correlation coefficients for the CHC factors reflect the total variance accounted for by that variable when it is included in a prediction equation in addition to the GIA. [b] Degrees of freedom reflects controlling for GIA. [c] Represents proportion of variance accounted for by factors entered jointly into the regression equation after controlling for the effects of the GIA. All values rounded to the nearest hundredth, may not sum to 100 due to rounding errors. * $p < .05$.

Table 21

Chi-Square Goodness of Fit Test Results Assessing the Significance of Obtained Variance Percentages across Levels of Schooling

	Broad Reading				Broad Mathematics				Broad Written Language			
Group Level	P_E	P_o	χ^2	*p*	P_E	P_o	χ^2	*p*	P_E	P_o	χ^2	*p*
K-5			2.69	.26			.04	.98			1.05	.59
General Factor	55	48			46	45			46	41		
CHC Factors	6	5			3	3			5	5		
Unpredictable [a]	39	47			51	52			49	54		
6-8			.20	.90			.37	.83			.22	.90
General Factor	55	55			46	49			46	46		
CHC Factors	6	7			3	3			5	6		
Unpredictable [a]	39	38			51	48			49	48		
9-12			3.55	.17			.43	.81			3.56	.17
General Factor	55	64			46	47			46	55		
CHC Factors	6	6			3	4			5	3		
Unpredictable [a]	39	30			51	49			49	42		

Note. P_E = expected percentages, P_o= observed percentages. *df* = 2. Expected values obtained from HMR analysis of total sample for each dependent variable. Variance percentages are R^2 and ΔR^2 multiplied by 100. CHC factor variance obtained using joint block entry procedures. [a] Combination of error and unaccounted for variance in the prediction model.

Table 22

Hierarchical Regression Analyses Predicting Broad Reading across Three Levels of Significant Gf-Gc Inter-Factor Variability

Discrepancy Level	Non-Significant (*n* = 1,529)[a]				15 Point (*n* = 530)[a]			
Predictor	R^2	95% CI	Δ R^2	Inc.[c]	R^2	95% CI	Δ R^2	Inc.[c]
GIA	.55*	.52-.58	-	55%	.53*	.48-.59	-	53%
CHC Factors (*df* = 7)[b]	.59*	.56-.62	.04	4%	.64*	.59-.69	.11	11%
Gc, Gf, Ga, Gv, Glr								
Gsm, Gs								

Discrepancy Level	23 Point (*n* = 190)[a]				30 Point (*n* = 65)[a]			
Predictor	R^2	95% CI	Δ R^2	Inc.[c]	R^2	95% CI	Δ R^2	Inc.[c]
GIA	.60*	.51-.69	-	60%	.56*	.41-.72	-	56%
CHC Factors (*df* = 7)[b]	.70*	.64-.77	.11	11%	-	-	-	**33%**
Crystallized Ability					.69*	.57-.81	.12	12%
Fluid Reasoning					.70*	.58-.82	.14	14%
Auditory Processing					.61*	.47-.75	.05	5%
Visual Processing					.57	.42-.72	.01	1%
Long-Term Retrieval					.57	.41-.72	.00	0%
Short-Term Memory					.56	.41-.72	.00	0%
Processing Speed					.57	.42-.72	.01	1%

Note. Gc = Crystallized Ability, Gf = Fluid Reasoning, Ga = Auditory Processing, Gv = Visual Processing, Glr = Long-Term Retrieval, Gsm = Short-Term Memory, Gs = Processing Speed. [a] Subsample utilizing listwise deletion. Squared multiple correlation coefficients for the CHC factors reflect the total variance accounted for by that variable when it is included in a prediction equation in addition to the GIA. [b] Degrees of freedom reflects controlling for GIA. [c] Represents proportion of variance accounted for by factors entered jointly into the regression equation after controlling for the effects of the GIA. **Individual percentages for each of the seven CHC factors were summed for the 30 point group**. All values rounded to the nearest hundredth, may not sum to 100 due to rounding errors. * $p < .05$.

Table 23

Hierarchical Regression Analyses Predicting Broad Mathematics across Three Levels of Significant Gf-Gc Inter-Factor Variability

Discrepancy Level	Non-Significant (*n* = 1,561)[a]				15 Point (*n* = 545)[a]			
Predictor	R^2	95% CI	ΔR^2	Inc.[c]	R^2	95% CI	ΔR^2	Inc.[c]
GIA	.46*	.42-.50	-	46%	.46*	.40-.52	-	46%
CHC Factors (*df* = 7)[b]	.49*	.45-.52	.03	3%	.49*	.43-.55	.03	3%
Gc, Gf, Ga, Gv, Glr Gsm, Gs								

Discrepancy Level	23 Point (*n* = 196)[a]				30 Point (*n* = 67)[a]			
Predictor	R^2	95% CI	ΔR^2	Inc.[c]	R^2	95% CI	ΔR^2	Inc.[c]
GIA	.48*	.38-.58	-	48%	.39*	.21-.56	-	39%
CHC Factors (*df* = 7)[b]	.53*	.44-.62	.05	5%	-	-	-	**9%**
Crystallized Ability					.41	.23-.58	.02	2%
Fluid Reasoning					.40	.23-.58	.02	2%
Auditory Processing					.40	.22-.57	.01	1%
Visual Processing					.39	.21-.56	.00	0%
Long-Term Retrieval					.40	.22-.57	.01	1%
Short-Term Memory					.39	.21-.56	.00	0%
Processing Speed					.42	.25-.59	.03	3%

Note. Gc = Crystallized Ability, Gf = Fluid Reasoning, Ga = Auditory Processing, Gv = Visual Processing, Glr = Long-Term Retrieval, Gsm = Short-Term Memory, Gs = Processing Speed. [a] Subsample utilizing listwise deletion. Squared multiple correlation coefficients for the CHC factors reflect the total variance accounted for by that variable when it is included in a prediction equation in addition to the GIA. [b] Degrees of freedom reflects controlling for GIA. [c] Represents proportion of variance accounted for by factors entered jointly into the regression equation after controlling for the effects of the GIA. **Individual percentages for each of the seven CHC factors were summed for the 30 point group**. All values rounded to the nearest hundredth, may not sum to 100 due to rounding errors. * $p < .05$.

Table 24

Hierarchical Regression Analyses Predicting Broad Written Language across Three Levels of Significant Gf-Gc Inter-Factor Variability

Discrepancy Level	Non-Significant (*n* = 1,528)[a]				15 Point (*n* = 534)[a]			
Predictor	R^2	95% CI	Δ R^2	Inc.[c]	R^2	95% CI	Δ R^2	Inc.[c]
GIA	.46*	.42-.50	-	46%	.45*	.39-.52	-	45%
CHC Factors (*df* = 7)[b]	.50*	.47-.54	.04	4%	.53*	.47-.59	.08	8%
Gc, Gf, Ga, Gv, Glr								
Gsm, Gs								

Discrepancy Level	23 Point (*n* = 191)[a]				30 Point (*n* = 68)[a]			
Predictor	R^2	95% CI	Δ R^2	Inc.[c]	R^2	95% CI	Δ R^2	Inc.[c]
GIA	.49*	.39-.59	-	49%	.44*	.27-.61	-	44%
CHC Factors (*df* = 7)[b]	.57*	.48-.65	.08	8%	-	-	-	**11%**
Crystallized Ability					.48*	.32-.64	.04	4%
Fluid Reasoning					.49*	.33-.65	.05	5%
Auditory Processing					.45	.28-.62	.01	1%
Visual Processing					.44	.27-.61	.00	0%
Long-Term Retrieval					.45	.28-.62	.01	1%
Short-Term Memory					.44	.27-.61	.00	0%
Processing Speed					.44	.27-.61	.00	0%

Note. Gc = Crystallized Ability, Gf = Fluid Reasoning, Ga = Auditory Processing, Gv = Visual Processing, Glr = Long-Term Retrieval, Gsm = Short-Term Memory, Gs = Processing Speed. [a] Subsample utilizing listwise deletion. Squared multiple correlation coefficients for the CHC factors reflect the total variance accounted for by that variable when it is included in a prediction equation in addition to the GIA. [b] Degrees of freedom reflects controlling for GIA. [c] Represents proportion of variance accounted for by factors entered jointly into the regression equation after controlling for the effects of the GIA. **Individual percentages for each of the seven CHC factors were summed for the 30 point group**. All values rounded to the nearest hundredth, may not sum to 100 due to rounding errors. * $p < .05$.

Table 25

Chi-Square Goodness of Fit Test Results Assessing the Significance of Obtained Variance Percentages across Three Gf-Gc Discrepancy Levels

	Broad Reading				Broad Mathematics				Broad Written Language			
Discrepancy Level	P_E	P_o	χ^2	*p*	P_E	P_o	χ^2	*p*	P_E	P_o	χ^2	*p*
15 Point			12.93	.00			.00	1.00			4.20	.12
General Factor	55	53			46	46			46	45		
CHC Factors	4	11			3	3			4	8		
Unpredictable [a]	41	36			51	51			50	47		
23 Point			16.22	.00			1.73	.42			5.18	.08
General Factor	55	60			46	48			46	49		
CHC Factors	4	11			3	5			5	8		
Unpredictable [a]	41	29			51	47			50	43		
30 Point			233.22	.00			15.17	.00			12.84	.00
General Factor	55	56			46	39			46	44		
CHC Factors	4	33			3	9			4	11		
Unpredictable [a]	41	11			51	52			50	45		

Note. P_E = expected percentages, P_o= observed percentages. *df* = 2. Expected values obtained from HMR analysis of less than 15 point Gf-Gc subsample for each dependent variable. Variance percentages are R^2 and ΔR^2 multiplied by 100. CHC factor variance obtained using joint block entry procedures, with the exception of the 30 point group. 30 point group CHC factor estimate is sum of variance accounted for by each factor entered individually in the second block of the regression equation. [a] Combination of error and unaccounted for variance in the prediction model.

Table 26

Hierarchical Regression Analyses Assessing the Impact of SLODR on the Predictive Validity of WJ-COG Variables across Differential Levels of General Ability

Broad Reading	Below Average (n = 169)[a]				Average (n = 1,678)[a]				Above Average (n = 191)[a]			
Predictor	R^2	95% CI	Δ R^2	Inc.[c]	R^2	95% CI	Δ R^2	Inc.[c]	R^2	95% CI	Δ R^2	Inc.[c]
GIA	.30*	.19-.42	-	30%	.33*	.30-.37	-	33%	.20*	.07-.33	-	20%
CHC Factors (*df* = 7)[b]	.41*	.30-.52	.11	11%	.41*	.37-.45	.08	8%	.35*	.25-.45	.15	15%
Gc, Gf, Ga, Gv, Glr												
Gsm, Gs												

Broad Mathematics	Below Average (n = 180)[a]				Average (n = 1,715)[a]				Above Average (n = 190)[a]			
Predictor	R^2	95% CI	Δ R^2	Inc.[c]	R^2	95% CI	Δ R^2	Inc.[c]	R^2	95% CI	Δ R^2	Inc.[c]
GIA	.24*	.13-.37	-	24%	.27*	.24-.31	-	27%	.13*	.04-.21	-	13%
CHC Factors (*df* = 7)[b]	.34*	.24-.45	.10	10%	.31*	.27-.35	.04	4%	.17	.07-.26	.04	4%
Crystallized Ability												
Gc, Gf, Ga, Gv, Glr												
Gsm, Gs												

Broad Written Language	Below Average (n = 173)[a]				Average (n = 1,682)[a]				Above Average (n = 188)[a]			
Predictor	R^2	95% CI	Δ R^2	Inc.[c]	R^2	95% CI	Δ R^2	Inc.[c]	R^2	95% CI	Δ R^2	Inc.[c]
GIA	.32*	.20-.43	-	32%	.27*	.23-.30	-	27%	.11*	.00-.21	-	11%
CHC Factors (*df* = 7)[b]	.42*	.31-.53	.10	10%	.32*	.28-.36	.06	6%	.24*	.14-.34	.14	14%
Gc, Gf, Ga, Gv, Glr												
Gsm, Gs												

Note. Gc = Crystallized Ability, Gf = Fluid Reasoning, Ga = Auditory Processing, Gv = Visual Processing, Glr = Long-Term Retrieval, Gsm = Short-Term Memory, Gs = Processing Speed. [a] Subsample utilizing listwise deletion. Squared multiple correlation coefficients for the CHC factors reflect the total variance accounted for by that variable when it is included in a prediction equation in addition to the GIA. [b] Degrees of freedom reflects controlling for GIA. [c] Represents proportion of variance accounted for by factors entered jointly into the regression equation after controlling for the effects of the GIA. All values rounded to the nearest hundredth, may not sum to 100 due to rounding errors. * $p < .05$.

Table 27

Chi-Square Goodness of Fit Test Results Assessing the Significance of the Impact of SLODR on Obtained Variance Percentages across Levels of General Ability

	Broad Reading				Broad Mathematics				Broad Written Language			
GIA Group Level	P_E	P_o	χ^2	*p*	P_E	P_o	χ^2	*p*	P_E	P_o	χ^2	*p*
Below Average			1.40	.50			9.46	.00			4.80	.09
General Factor	33	30			27	24			27	32		
CHC Factors	8	11			4	10			6	10		
Unpredictable [a]	59	59			69	66			67	58		
Above Average			11.86	.00			10.10	.00			21.10	.00
General Factor	33	20			27	13			27	11		
CHC Factors	8	15			4	4			6	14		
Unpredictable [a]	59	65			69	83			67	75		

Note. P_E = expected percentages, P_o= observed percentages. *df* = 2. Expected values obtained from HMR analysis of average ability subsample for each dependent variable. Variance percentages are R^2 and ΔR^2 multiplied by 100. CHC factor variance obtained using joint block entry procedures. [a] Combination of error and unaccounted for variance in the prediction model.

Table 28

Synthesis of Un-weighted Hierarchical Multiple Regression Model Estimates for the Prediction of Reading, Math, and Writing Outcomes, Utilizing a Fixed Effects Meta-Analytic Format

				Reading	Math	Writing
Study	IQ Test	*N*	Predictor	$R^2/\Delta R^2$	$R^2/\Delta R^2$	$R^2/\Delta R^2$
Glutting, Youngstrom, Ward, Ward, & Hale (1997)	WISC-III	283	Full-Scale	.42	.56	.32
			Factor Scores (*df* = 4)[b]	.08	.10	.16
			Full-Scale	.42	.56	.32
			VIQ-PIQ (*df* = 2)[b]	.04	.02	.03
Youngstrom, Kogos, & Glutting (1999)	DAS	1,185	Full-Scale	.36	.35	.27
			Factor Scores (*df* = 3)[b]	.04*	.02*	.03*
Glutting, Watkins, Konold, & McDermott (2006)	WISC-IV	498	Full-Scale	.60	.60	-
			Factor Scores (*df* = 4)[b]	.02	.00	-
Canivez (2013a)	WAIS-IV	93	Full-Scale	.58	.71	.42
			Factor Scores (*df* = 4)[b]	.10	.05	.01
		59	Full-Scale	.40	.53	.39
			Factor Score (*df* = 4)[b]	.08	.05	.12
Canivez (2011)	CAS	1,600	Full-Scale	.41	.44	.42
			Factor Scores (*df* = 4)[b]	.02	.02	.01
McGill & Busse (2012)	KABC-II	2,520	Full-Scale	.55	.50	.43
			Factor Scores (*df* = 5)[b]	.04	.01	.02
[c]McGill (2013)	WJ-COG	2,059[a]	Full Scale	.55	.46	.46
			Factor Scores (*df* = 7)[b]	.06	.03	.05

Note. [a] Smallest listwise sample across the three independent variables in the HMR analyses. *Indicates subtest outcome measure. Values rounded to the nearest hundredth. [c] Current study not included in calculation of grand effect sizes.

Table 29

Chi-Square Goodness of Fit Test Results Comparing Achievement Variance Accounted for by the CHC Model on the WJ-COG with Aggregated Estimates Obtained in HMR Analyses of Other Intelligence Test Measures

	Reading				Mathematics				Writing			
Variance	P_E	P_o	χ^2	*p*	P_E	P_o	χ^2	*p*	P_E	P_o	χ^2	*p*
			3.71	.16			.00	1.00			3.33	.19
Full-Scale	48	55			46	46			39	46		
Factor Scores	4	6			3	3			3	5		
Unpredictable [a]	48	39			51	51			58	51		

Note. P_E = expected percentages, P_o= observed percentages. *df* = 2. Expected values are grand R^2 values multiplied by 100 using weighted values tabulated from Table 28. Values synthesized using a fixed effects meta-analytic model. Variance percentages are R^2 and ΔR^2 multiplied by 100. Factor score variance obtained using joint block entry procedures. [a] Combination of error and unaccounted for variance in the prediction model.

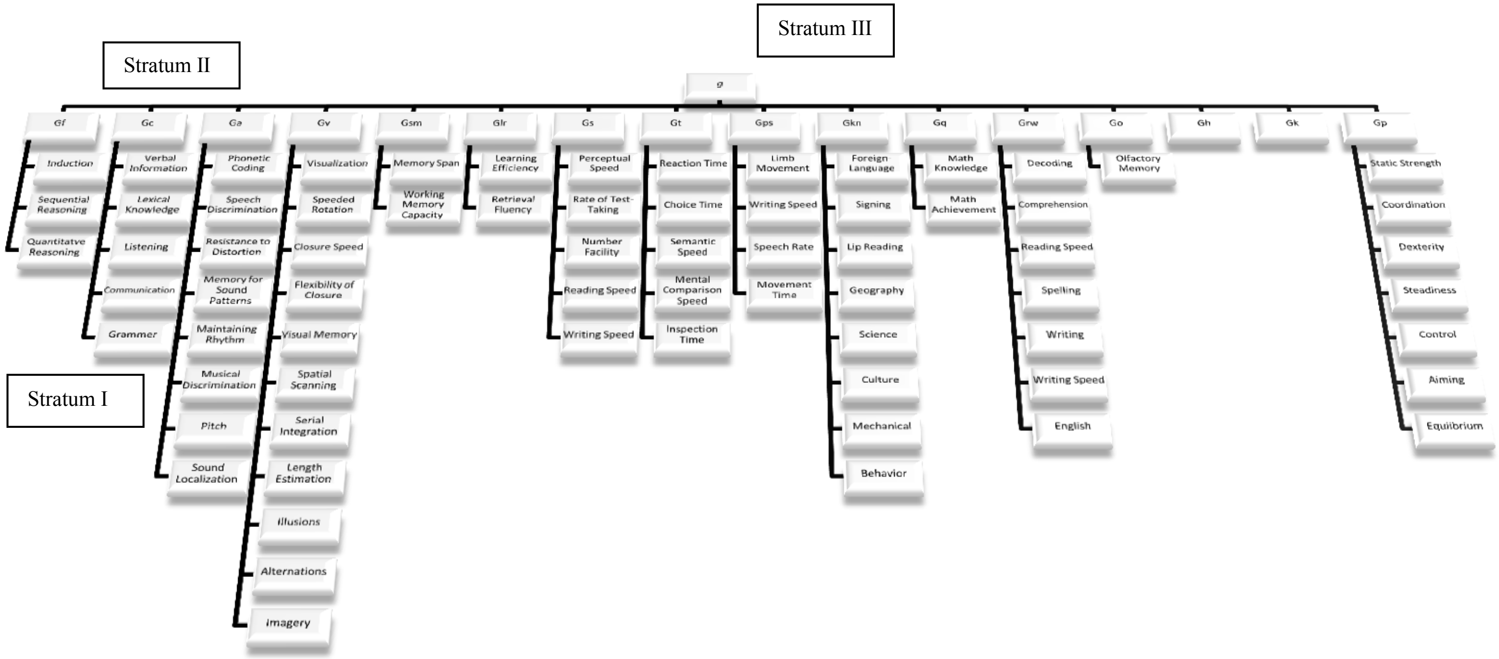


Figure 1. The most recently updated version of the Cattell-Horn-Carroll (CHC) Model of Human Cognitive Abilities (Schneider & McGrew 2012). g = general intelligence, Gf = fluid intelligence/reasoning, Gq = quantitative reasoning, Gc = crystallized intelligence/ability, Grw = general reading and writing ability, Gsm = short-term memory, Gv = visual processing, Ga = auditory processing, Glr = long-term memory and retrieval, Gs = processing speed, Gt = decision speed, Gps = psychomotor speed, Gkn = domain specific knowledge, Go = olfactory abilities, Gh = tactile abilities, Gk = kinesthetic abilities, Gp = psychomotor abilities.

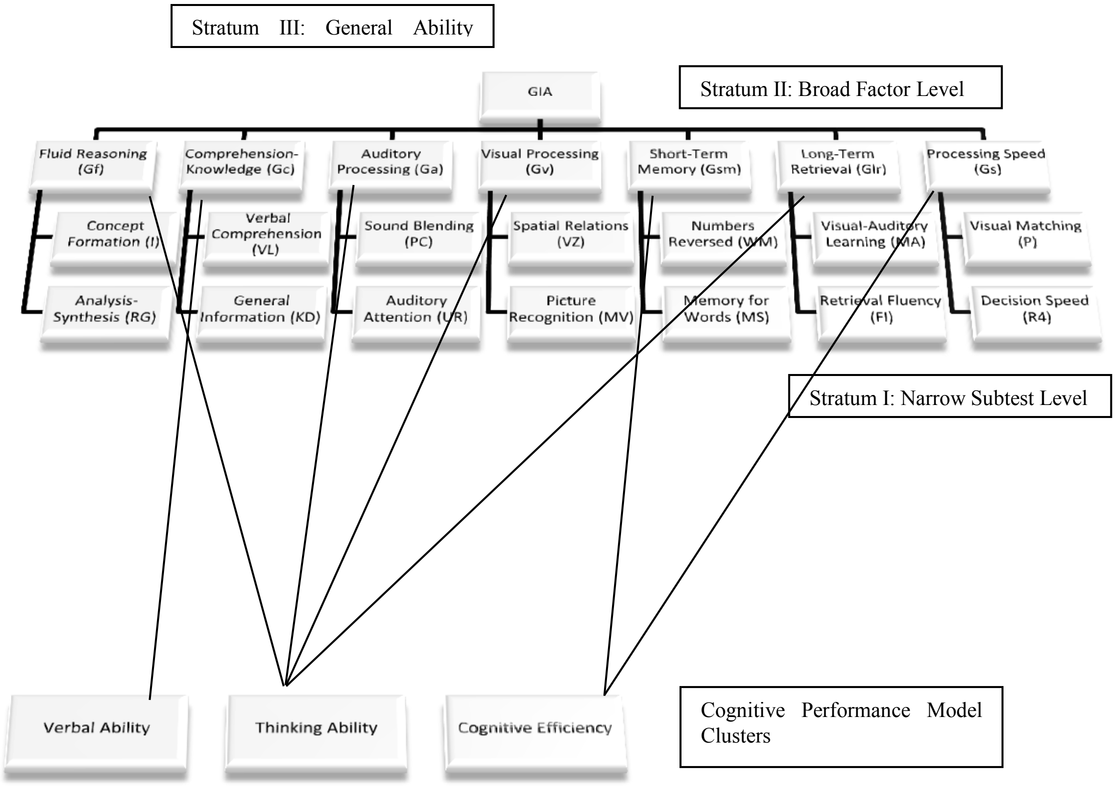


Figure 2. Interpretative structure and organization of the WJ-COG. *Note*. Additional CHC stratum II broad abilities are incorporated into the design of the WJ-ACH which is not represented here. I = induction, RG = sequential reasoning, VL = lexical knowledge, KD = verbal information, PC = phonetic coding, UR = resistance to distortion, VZ = visualization, MV = visual memory, WM = working memory capacity, MS = memory span, MA = associative memory, FI = ideational fluency, P = perceptual speed, R4 = semantic processing speed.

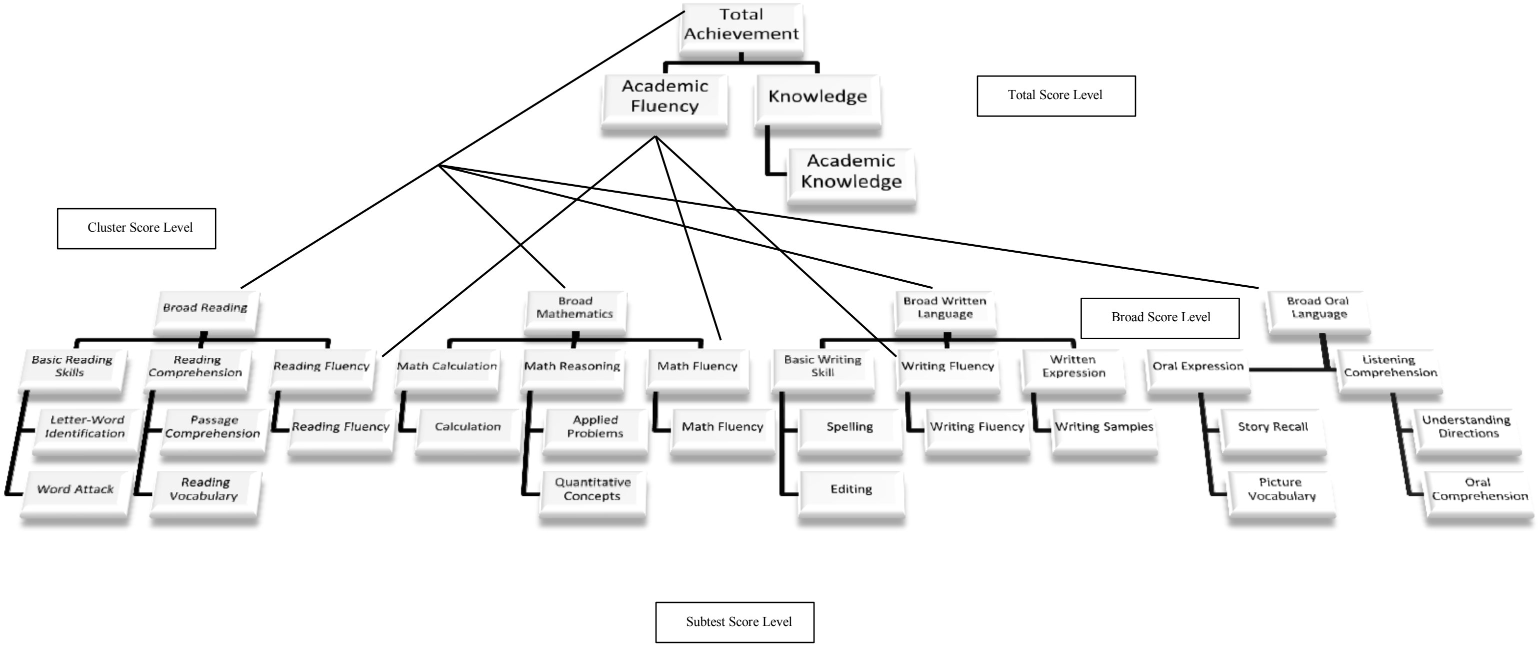


Figure 3. Structure and organization of the WJ-ACH. *Note*. Although the WJ-ACH is designed to measure additional CHC broad abilities (e.g., Gq and Grw), they are not the focus of analysis as this study is concerned with assessing the incremental validity of cognitive broad abilities.

Research Question	Hypothesis	IV	DV	Method of Analysis
Q1: Do the CHC broad ability factors on the WJ-COG provide *statistically* significant incremental prediction of achievement outcomes after the effects of the general factor have been controlled?	**H1:** Yes, it is expected that the CHC broad ability factors will predict *statistically* significant achievement variance on the WJ-COG beyond that already accounted for by the GIA.	First Block: WJ-COG GIA Second Block: WJ-COG Gc, Gf, Glr, Gsm, Ga, Gv, and Gs factor scores (jointly and individually entered)	WJ-ACH Reading Comprehension, Basic Reading Skills, Math Reasoning, Math Calculation Skills, Written Expression, Listening Comprehension, and Oral Expression composite standard scores	Hierarchical Multiple Regression Analysis of variance (ANOVA) tests of significance and *F* test statistic
Q2: Do the CHC broad ability factors on the WJ-COG provide *clinically* significant incremental prediction of achievement outcomes after the effects of the general factor have been controlled?	**H2:** No, it is not expected that the CHC broad ability factors will predict *clinically* significant achievement variance on the WJ-COG beyond that already accounted for by the GIA.	First Block: WJ-COG GIA Second Block: WJ-COG Gc, Gf, Glr, Gsm, Ga, Gv, and Gs factor scores (jointly and individually entered)	WJ-ACH Reading Comprehension, Basic Reading Skills, Math Reasoning, Math Calculation Skills, Written Expression, Listening Comprehension, and Oral Expression composite standard scores	Hierarchical Multiple Regression R^2 and ΔR^2 coefficients as effect size estimates with corresponding 95% confidence intervals
Q3: Is the predictive validity of the CHC factors on the WJ-COG invariant across	**H3:** Yes, it is expected that the predictive validity of the CHC factors will not be	First Block: WJ-COG GIA Second Block: WJ-COG Gc, Gf,	WJ-ACH Broad Reading, Broad Mathematics, and Broad Written Language scores	Hierarchical Multiple Regression Chi-square goodness of fit test using R^2 and

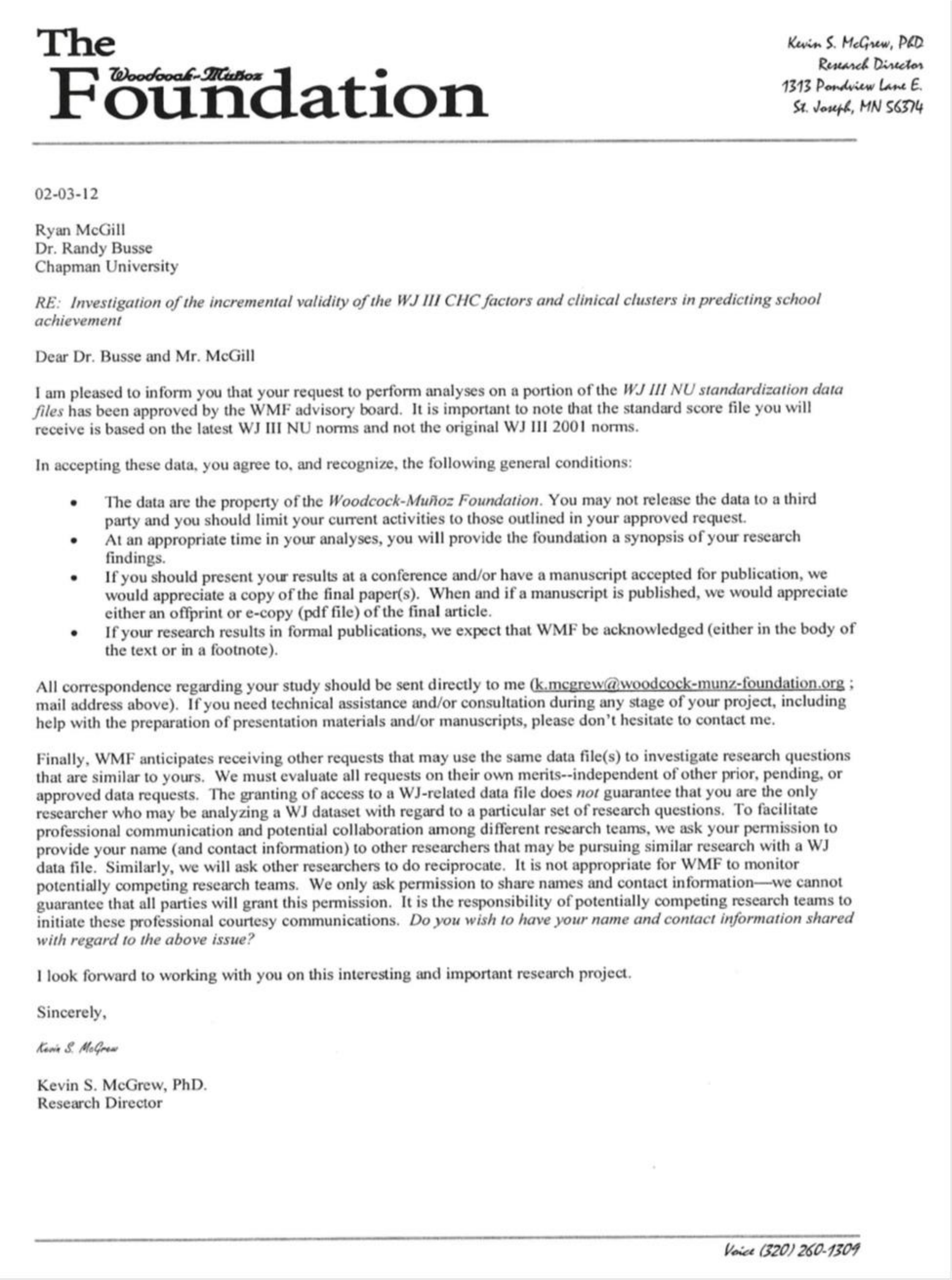
different levels of schooling?	invariant across levels of schooling.	Glr, Gsm, Ga, Gv, and Gs factor scores (jointly entered)		ΔR^2 coefficients as percentages and the variance percentages obtained from question 2 as parameter estimates
Q4: Is the predictive validity of *g* attenuated by significant levels of inter-factor variability on the WJ-COG?	**H4:** No, it is not expected that the predictive validity of *g* will be impacted by significant levels of inter-factor scatter.	First Block: WJ-COG GIA Second Block: WJ-COG Gc, Gf, Glr, Gsm, Ga, Gv, and Gs factor scores (individually entered)	WJ-ACH Broad Reading, Broad Mathematics, and Broad Written Language standard scores	Hierarchical Multiple Regression Chi-square goodness of fit test using R^2 and ΔR^2 coefficients as percentages and the variance percentages obtained from a control group (less than 15 point Gf-Gc difference) for parameter estimates
Q5: Does SLODR impact the predictive validity of the general factor on the WJ-III COG?	**H5:** Yes, it is expected that the predictive validity of the WJ-COG GIA will be impacted by the presence of SLODR.	First Block: WJ-COG GIA Second Block: WJ-COG Gc, Gf, Glr, Gsm, Ga, Gv, and Gs factor scores (jointly and entered)	WJ-ACH Broad Reading, Broad Mathematics, and Broad Written Language scores	Hierarchical Multiple Regression Chi-square goodness of fit test using R^2 and ΔR^2 coefficients as percentages and the variance percentages obtained from the average group as parameter estimates
Q6: Does the use of a differential weighting scheme	**H6:** Yes, The predictive utility of the CHC factor	First Block: General Factor Estimate	Various reading, mathematics, and written language composites from norm-referenced standardized	Interpretation of R^2 and ΔR^2 coefficients as effect sizes within a

enhance the validity of the WJ-COG factor structure in predicting norm-referenced reading, math, and writing outcomes when compared to estimates that have been obtained from other intelligence tests using similar methods of variance partitioning on commercial standardization samples?	structure on the WJ-COG will be consistent with model estimates that have been obtained from HMR analyses of other intelligence tests.	Second Block: Broad Factor Estimates (jointly entered)	achievement tests	fixed effects meta-analytic format for comparisons across models. Chi-square goodness of fit test comparisons.

Figure 4. Matrix of research questions, hypothesis, variables, and methods of analysis.

Appendix A

Woodcock- Muñoz Foundation Data Release Agreement



The Woodcock-Muñoz Foundation

Kevin S. McGrew, PhD
Research Director
1313 Pondview Lane E.
St. Joseph, MN 56374

02-03-12

Ryan McGill
Dr. Randy Busse
Chapman University

RE: Investigation of the incremental validity of the WJ III CHC factors and clinical clusters in predicting school achievement

Dear Dr. Busse and Mr. McGill

I am pleased to inform you that your request to perform analyses on a portion of the *WJ III NU standardization data files* has been approved by the WMF advisory board. It is important to note that the standard score file you will receive is based on the latest WJ III NU norms and not the original WJ III 2001 norms.

In accepting these data, you agree to, and recognize, the following general conditions:

- The data are the property of the *Woodcock-Muñoz Foundation*. You may not release the data to a third party and you should limit your current activities to those outlined in your approved request.
- At an appropriate time in your analyses, you will provide the foundation a synopsis of your research findings.
- If you should present your results at a conference and/or have a manuscript accepted for publication, we would appreciate a copy of the final paper(s). When and if a manuscript is published, we would appreciate either an offprint or e-copy (pdf file) of the final article.
- If your research results in formal publications, we expect that WMF be acknowledged (either in the body of the text or in a footnote).

All correspondence regarding your study should be sent directly to me (k.mcgrew@woodcock-munz-foundation.org ; mail address above). If you need technical assistance and/or consultation during any stage of your project, including help with the preparation of presentation materials and/or manuscripts, please don't hesitate to contact me.

Finally, WMF anticipates receiving other requests that may use the same data file(s) to investigate research questions that are similar to yours. We must evaluate all requests on their own merits--independent of other prior, pending, or approved data requests. The granting of access to a WJ-related data file does *not* guarantee that you are the only researcher who may be analyzing a WJ dataset with regard to a particular set of research questions. To facilitate professional communication and potential collaboration among different research teams, we ask your permission to provide your name (and contact information) to other researchers that may be pursuing similar research with a WJ data file. Similarly, we will ask other researchers to do reciprocate. It is not appropriate for WMF to monitor potentially competing research teams. We only ask permission to share names and contact information—we cannot guarantee that all parties will grant this permission. It is the responsibility of potentially competing research teams to initiate these professional courtesy communications. *Do you wish to have your name and contact information shared with regard to the above issue?*

I look forward to working with you on this interesting and important research project.

Sincerely,

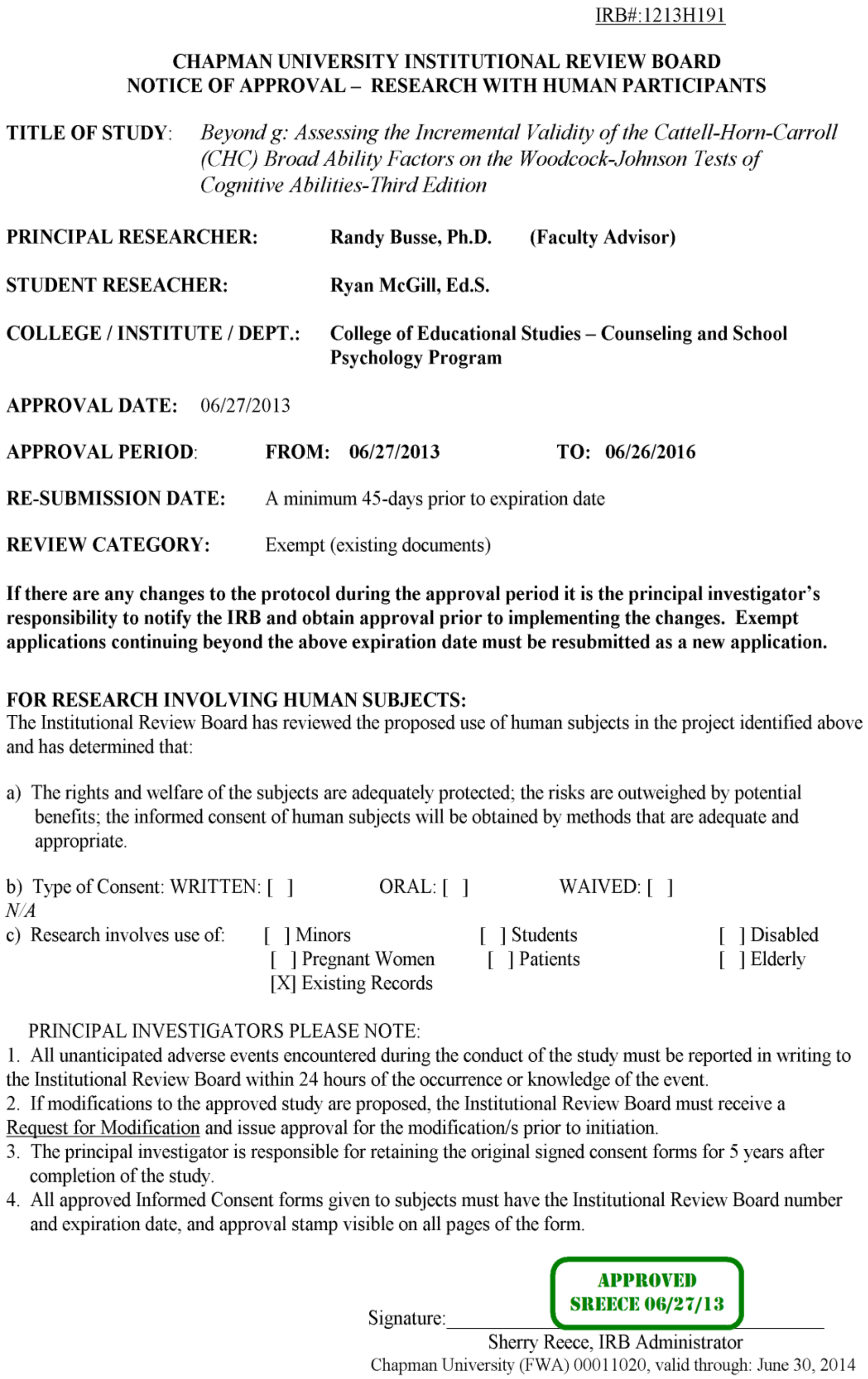
Kevin S. McGrew

Kevin S. McGrew, PhD.
Research Director

Voice (320) 260-1309

Appendix B

Chapman University Institutional Review Board Approval Notice



IRB#:1213H191

CHAPMAN UNIVERSITY INSTITUTIONAL REVIEW BOARD
NOTICE OF APPROVAL – RESEARCH WITH HUMAN PARTICIPANTS

TITLE OF STUDY: *Beyond g: Assessing the Incremental Validity of the Cattell-Horn-Carroll (CHC) Broad Ability Factors on the Woodcock-Johnson Tests of Cognitive Abilities-Third Edition*

PRINCIPAL RESEARCHER: Randy Busse, Ph.D. (Faculty Advisor)

STUDENT RESEACHER: Ryan McGill, Ed.S.

COLLEGE / INSTITUTE / DEPT.: College of Educational Studies – Counseling and School Psychology Program

APPROVAL DATE: 06/27/2013

APPROVAL PERIOD: **FROM: 06/27/2013 TO: 06/26/2016**

RE-SUBMISSION DATE: A minimum 45-days prior to expiration date

REVIEW CATEGORY: Exempt (existing documents)

If there are any changes to the protocol during the approval period it is the principal investigator's responsibility to notify the IRB and obtain approval prior to implementing the changes. Exempt applications continuing beyond the above expiration date must be resubmitted as a new application.

FOR RESEARCH INVOLVING HUMAN SUBJECTS:
The Institutional Review Board has reviewed the proposed use of human subjects in the project identified above and has determined that:

a) The rights and welfare of the subjects are adequately protected; the risks are outweighed by potential benefits; the informed consent of human subjects will be obtained by methods that are adequate and appropriate.

b) Type of Consent: WRITTEN: [ ] ORAL: [ ] WAIVED: [ ]
N/A

c) Research involves use of: [ ] Minors [ ] Students [ ] Disabled
[ ] Pregnant Women [ ] Patients [ ] Elderly
[X] Existing Records

PRINCIPAL INVESTIGATORS PLEASE NOTE:

1. All unanticipated adverse events encountered during the conduct of the study must be reported in writing to the Institutional Review Board within 24 hours of the occurrence or knowledge of the event.
2. If modifications to the approved study are proposed, the Institutional Review Board must receive a Request for Modification and issue approval for the modification/s prior to initiation.
3. The principal investigator is responsible for retaining the original signed consent forms for 5 years after completion of the study.
4. All approved Informed Consent forms given to subjects must have the Institutional Review Board number and expiration date, and approval stamp visible on all pages of the form.

APPROVED
SREECE 06/27/13

Signature:____________________________________
Sherry Reece, IRB Administrator
Chapman University (FWA) 00011020, valid through: June 30, 2014