A COMPARISON OF CONFIRMATORY FACTOR ANALYSIS AND TASK ANALYSIS OF FLUID INTELLIGENCE COGNITIVE SUBTESTS

A dissertation presented to the Faculty of the Graduate School at the University of Missouri

In Partial Fullfillment of the Requirements for the Degree

Doctor of Philosophy

by
JASON R. PARKIN

Dr. Craig L. Frisby, Dissertation Supervisor

MAY 2010

The undersigned, appointed by the dean of the Graduate School, have examined the dissertation entitled

A COMPARISON OF CONFIRMATORY FACTOR ANALYSIS AND TASK ANALYSIS OF FLUID INTELLIGENCE COGNITIVE SUBTESTS

presented by Jason R. Parkin,

a candidate for the degree of doctor of philosophy,

and hereby certify that, in their opinion, it is worthy of acceptance.

Professor Craig Frisby

Professor James Koller

Professor Cheryl Offutt

Professor Steven Osterlind

Professor Erica Lembke

DEDICATION

I dedicate this project to my family – my parents Richard and Katherine, my brother Erik, my sister Kristen, and my fiancé Sharon. Their support, encouragement and good humor were important in so many ways through out these years.

Thank you so much!

ACKNOWLEDGMENTS

Studying school psychology at the University of Missouri has been a great pleasure. As my time there draws to a close, there are several individuals I would like to acknowledge for contributing to my training and research experiences. Dr. Craig Frisby deserves a special thanks for both his guidance and patience during my studies and the conceptualization and execution of this research project. His good humor was especially important during the numerous drafts of this paper.

I also would like to acknowledge my committee members, Drs. Craig Frisby, James Koller, Cheryl Offutt, Steven Osterlind and Erica Lembke. Their unique expertise and perspectives, along with their support made this project something I can be proud of.

I would also like to thank Drs. Jim Koller and Karen Weston at the Center for the Advancement of Mental Health Practices in Schools (CAMHPS). My training with these individuals strongly influenced my conceptualization of school psychology practice and provided many great learning experiences and wonderful friends.

Lastly, I want to acknowledge the professionals at serving the public schools that participated in this project. Without their help and support, this study would not have been possible.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ii

LIST OF TABLES v

LIST OF ILLUSTRATIONS vi

ABSTRACT vii

Chapter

1. INTRODUCTION 1

   History of Cognitive Test Interpretation

   The Practice of Cross-Battery Assessment

   Cross-Battery Assessment Knowledge Basis

   Cross-Battery Assessment Guiding Principles

   Interpretation of Cross-Battery Assessments

   Purpose of Study

2. REVIEW OF LITERATURE 16

   History of Factor Analytic Theories of Intelligence and the Development of CHC Theory

   Applying CHC Theory to Major Cognitive Batteries through Factor Analysis

   Narrow Ability Measures

   Task Analytic Methods in Subtest Classification

   History of Task Analysis in Test Interpretation

   Task Analysis and Information Processing Theories

   Fluid Intelligence and its Task Demands

   Introduction to the Study

Hypotheses

3. METHODS 56

Approval by Human Subjects Committee

Participants

Instruments

Procedure

Data Analysis

4. RESULTS 70

Phase 1: Classification Based on Factor Analysis

Phase 2: Classification Based on Task Analysis

5. DISCUSSION 85

Summary of Results

Comparison of Findings

Implications of Findings

Limitations and Future Directions

Conclusions

APPENDIXES

1. Subtest Task Descriptions...95

2. Data Sheets, Informed Consent Forms and Recruitment Materials....................101

REFERENCES 109

VITA 122

LIST OF TABLES

Table Page

1. Card Sorting Sample Demographic Information 58

2. Normality Testing 71

3. Descriptive Statistics and Correlations for Gf Subtests for n = 63 Participants 73

4. Multicollinearity diagnositics with six Gf subtests..73

5. Multicollinearity diagnostics with four subtests...75

6. Fit Statistics Comparing One-Factor and Two-Factor Models....................................78

7. Subtest Cluster Membership by Cluster Solution..79

8. Three Cluster Solution 80

9. Four Cluster Solution..81

10. Five Cluster Solution 82

11. Six Cluster Solution..82

12. Seven Cluster Solution...83

LIST OF ILLUSTRATIONS

Illustration Page

1. Hypothesized one-factor model for confirmatory factor analysis 54

2. Hypothesized three-factor model for confirmatory factor analysis 54

3. One-factor model with four subtests, including factor loadings..................................77

4. Two-factor model with four subtests, including factor loadings 77

A COMPARISON OF CONFIRMATORY FACTOR ANALYSIS AND TASK ANALYSIS OF FLUID INTELLIGENCE COGNITIVE SUBTESTS

Jason R. Parkin

Dr. Craig Frisby, Dissertation Supervisor

ABSTRACT

Cross-battery assessment relies on the classification of cognitive subtests into the Cattell-Horn-Carroll (CHC) theory's broad and narrow ability definitions. Generally, broad ability classifications have used ability data analyzed through factor analytic methods, while narrow ability classifications have used data about subtest task demands. The purpose of this investigation is to determine whether subtest similarity judgments based on task demands data, and judgments based on ability measurement provide similar results. It includes two studies. First, middle school students (N = 63) completed six target fluid reasoning subtests that were subjected to confirmatory factor analyses to analysis subtest similarities. Second, school psychology practitioners (N = 32) sorted subtest descriptions into similarity groups. Their judgments were analyzed with multiple non-hierarchical cluster analyses. Results partially confirmed that the six target subtests were classified similarly using both data types, though need to be interpreted cautiously due to limitations. Implications for assessment practices are discussed.

CHAPTER ONE: INTRODUCTION

IQ (intelligence) testing enjoys popular and widespread use among school psychologists (Pfeiffer, Reddy, Kletzel, Schmeizer & Boyer, 2000). Each year, over one million IQ tests are individually administered to children in schools (Macmann & Barnett, 1997). By median estimate, practitioners administer two to three IQ tests each week (Pfeiffer et al., 2000). Although IQ testing has always been a constant part of school psychological practice, the way clinicians have interpreted IQ tests has evolved throughout the history of testing. As part of that evolution, Flanagan and colleagues introduced a "cross-battery" method of test interpretation in the late 1990's (Flanagan, Ortiz & Alfonso, 2007).

The term "cross-battery assessment" describes the practice of supplementing cognitive test batteries by using subtests from other cognitive tests in the context of a single individual assessment (Flanagan & McGrew, 1997; Flanagan, McGrew & Ortiz, 2000; Flanagan, Ortiz & Alfanso, 2007; McGrew & Flanagan, 1998). Its application rests on a synthesized version of John Carroll's Three-Stratum theory (Carroll,1993) and Catell-Horn's Gf-Gc theory (Horn & Noll, 1997). The resulting intergration, termed the Cattell-Horn-Carroll theory (CHC; see McGrew, 2009), provides a standard nomenclature to describe a spectrum of human cognitive abilities for potential application to all cognitive test batteries. Use of CHC theory and cross-battery assessment in clinical practice represents an effort to maximize how well a test represents the specturm of human cognitive ability to improve construct representation (Cohen & Swerdlik, 2001; Sattler, 2001).

Whereas Spearman's (1904) original theory of human intelligence focused on "*g"* as a critical construct of measurement, CHC theory describes the construct of intelligence as a hierarchy of abilities. Spearman's "g" reflects the most fundamental source of individual differences in mental ability, and the broadest level of ability (Stratum III). It subsumes 9 to 10 more specific, broad cognitive (Stratum II) abilities. These broad abilities subsume even more specific narrow abilities (Stratum I). The hierarchical nature of CHC theory stresses that a narrow ability, such as length estimation (the ability to estimate visual lengths without using measurement tools), describes just one part of visual-spatial processing, a broad ability. Similarly, a broad ability like visual-spatial processing reflects just one part of general cognitive functioning.

The practice of cross-battery assessment rests on single battery and multiple-battery factor analyses that classify cognitive subtests into CHC's Stratum II ability definitions. However, these factor analyses have not adequately classified subtests into narrow, Stratum I abilities. Instead, subtests have been sorted into narrow abilities through use of a rational, expert consensus method (McGrew, 1997). Scholars have relied on this method because a study designed to partition any particular subtest into its specific ability components (ie. "g", broad abilities, and specific, narrow abilities) would be a major undertaking and require a very large sample on which a variety of subtests are factor analyzed (Phelps, McGrew, Knopik & Ford, 2005). Unlike factor analysis, expert consensus methods cannot rely upon the shared variance between ability scores to understand the ability(ies) a subtest may measure. Instead, these methods use experts' clinical judgments of subtests' task similiarities to classify subtests that may appear to measure similar constructs. Frisby and Parkin (2007) have suggested that rational methods of classifying subtests are associated with a number of limitations. Most importantly, these methods may confound ability and task demand variables. Accordingly, the classification of subtests into the narrowest of ability categories may represent a limitation of cross-battery assessment. The proposed study hopes to begin to address this limitation by comparing factor and task analyses of subtest classification.

History of Cognitive Test Interpretation

Throughout the history of testing, cross-battery assessment has consistently represented an effort to integrate modern theories of cognitive abilities with the applied practice of cognitive assessment. Across time, scholars have struggled to adequately define intelligence. In 1921, when the *Journal of Educational Psychology* published a symposium where invited experts debated the nature of intelligence, there was little agreement on an appropriate definition (Cohen & Swerdlik, 2002), a necessity for guiding IQ score interpretation. Due to the controversy and ambiguity associated with the nature of intelligence as a construct, E.G. Boring suggested that “intelligence is what the tests test” (Cohen & Swerdlik, 2002, p. 226). Such a circular definition suggests that past psychometric methods of intelligence testing may have lacked an adequate theoretical foundation for interpretation. Even more recently, researchers have noted that newer test batteries, such as the Wechsler Intelligence Scale for Children – Third Edition (WISC-III), “was not derived from a specific theoretical structure” (Watkins, Greenawalt & Marcell, 2002, p. 165; see also Esters, Ittenbach & Han, 1997). Other scholars have noted similar theoretical problems with the Wechsler Intelligence Scale for Children – Fourth Edition (WISC-IV; Keith et al., 2006; Wechsler, 2003a; Wechsler, 2003b).

Intelligence test interpretation throughout the decades. Perhaps as a result of the difficulty in creating a theoretical definition of intelligence, the interpretation of intelligence tests has appeared atheoretical throughout its history. Kamphaus, Petoskey, and Morgan (1997) describe the history of intelligence test interpretation in four stages, or waves. In the first wave, dated around the beginning of the 20th century, and marked by the creation of Binet's first IQ test, a cognitive score was used for the quantification and classification of individuals into categories of intellectual functioning. The second wave began around the later half of the 20th century, when IQ tests began to include subtests scores (Kamphaus, Petoskey & Morgan), clinical profile analysis defined test interpretation. With profile analysis, clinicians began to compare an individual's performance on one task relative to their performance on another task. Test interpretation's third wave began around the 1960s and 1970s when personal computers made factor analysis easier to perform. Kamphaus and colleagues (1997) stress that factor analytic studies demonstrated that many of the clinical methods of profile analysis used in the second wave of interpretation lacked empirical support. While this represented an important step forward, interpretation of IQ scores generally lacked a theoretical context to guide the process of assessment. It was not until the more modern fourth wave of IQ score interpretation that "theory and measurement science [were] intermingled" (p. 41). Kamphaus et al. (1997) labeled the fourth wave of test interpretation "Applying Theory to Intelligence Test Interpretation" (p. 43). These authors suggest that the fourth wave can be characterized by an integrative method of interpretation where IQ scores are interpreted in the context of other clinical findings, background information, and research results. They stress the importance of a priori clinical hypothesis that are confirmed or disconfirmed by test results.

Kamphaus et al.'s (1997) predictions for a future fifth wave of intelligence test interpretation focus on tests' content validity, which favors strong theoretical definitions of intelligence (Kamphaus, Petoskey & Morgan, 1997). Developments in Kamphaus et al.'s fifth wave can be considered somewhat similar to developments in the fourth wave of interpretation. Whereas in the fourth wave, clinicians created hypothesis prior to assessment in order to be disconfirmed by the results, in the fifth wave, test developers may begin to use modern theories of intelligence prior to the construction of newer tests. The initial development of the Woodcock-Johnson Tests of Cognitive Ability - Revised (Woodcock & Johnson, 1989) may represent such a focus with the application of Gf-Gc theory (Horn & Noll, 1997) in its development (Kamphaus et. al.). In general, the fifth wave's focus on content validity underscores the need to strengthen the link between theory and practice in intelligence assessment (e.g. Woodcock, 1998).

The Practice of Cross-Battery Assessment

Cross-battery assessment encourages clinicians to move past the Full Scale IQ (FSIQ) when assessing cognitive ability, a controversial practice (McDermott, Fantuzzo & Glutting, 1990; McDermott, Fantuzzo, Glutting, Watkins & Baggaley, 1992; Watkins, Glutting, & Youngstrom, 2002, 2003). McGrew and Flanagan (1997) suggest that a lack of adequate theory to guide test interpretation, and the poor content validity of many test batteries may be key reasons why many scholars and researchers have advocated for test interpretation to remain at the level of the FSIQ. Moving interpretation past the FSIQ involves profile analysis, or the interpretation of an examinees' patterns of test performance (Sattler, 2001). Intraindividual profile analysis (e.g. ipastive analysis) compares examinee performance on one set of tasks to his or her performance on another set of tasks. To calculate ipsative scores, clinicians subtract an examinee's performance on a subtest from the average of his performance on all subtests administered (McDermott et. al., 1992; Sattler, 2001). Alternatively, interindividual profile analysis compares examinee performance on a set of tasks to a normative group (Sattler).

When conducting intra and interindividual profile analyses, clinicians engage in the process of task analysis. Task analysis involves the comparison of functional task demands - mode of stimuli input and response output, required strategies and cognitive processes - from of a number of individual subtests in an effort to understand the relationship of peaks and valleys between those subtests. As Kaufman (1994) demonstrates, subtests may be grouped based upon their sensitivity to attention and concentration, motor demands, visual organization, lack of motor coordination, or any other kind of functional requirement. Perhaps because test interpretation may lack a comprehensive theoretical base, clinicians and researchers frequently analyze subtests' functional demands using rational or intuitive methods (Frisby & Parkin, 2007). As a result, task analyses may diverge "in accordance with each author's personal orientation" (Kaufman, 1994, p.49) rather than reflect theoretically based constructs validated through scientific research.

Some research has indicated that the practice of profile analysis adds little to the test interpretation process (McDermott, Fantuzzo & Glutting, 1990). For instance, McDermott and colleagues (1992) have demonstrated that intraindividual analysis of Wechsler Intelligence Scale for Children – Revised (WISC-R; Wechsler, 1974) subtest scores explains about a third of the variance in achievement data that comparing subtest performance to a normative group explains (e.g. normative analysis). These researchers also demonstrated that when regressed together on academic achievement, intraindividual analysis of cognitive scores (the difference between an examinee's subtest performance compared to the mean of all subtests) does not explain additional variance after accounting for variance explain by interindividual analysis of cognitive scores. Researchers have also documented similar findings using the Wechsler Intelligence Scale for Children – Third Edition (WISC-III; Wechsler, 1991) and the Differential Ability Scale (DAS; Elliot, 1990) test batteries (McDermott & Glutting, 1997).

Though interindividual analysis explains more variance than intraindividual analysis, some researchers also advocate against its use in clinical practice as well. Joseph Glutting and colleagues (Glutting, Watkins, Konold & McDermott, 2006; Kahana, Youngstrom, & Glutting, 2002; McDermott, Fantuzzo, & Glutting, 1990; Watkins, Glutting & Youngstrom, 2005; Youngstrom, Kogos, & Glutting, 1999; Watkins, Glutting & Lei, 2007) argue that a FSIQ is the best predictor of academic achievement and stress that currently there is no evidence that more specific abiltities can match the ability of *g* to predict a variety of achievement outcomes (Watkins, Glutting, & Youngstrom, 2002, 2003). They also emphasize that there is minimal difference between the amount of variance explained by a group of more specific cognitive abilities when compared to *"g."*

However, others argue that it may be premature to stop the search for more specific abilities that influence learning (Carroll, 1993; McGrew & Flanagan, 1997; Oh, Glutting, Watkins, Youngstrom, & McDermott, 2004). Many of these scholars suggest that negative research findings regarding the importance of specific cognitive abilities (in contrast to the single general factor "*g*") may be due to a lack of content validity in modern tests. Several cognitive batteries have lacked strong theoretical evidence (Kamphaus, Petoskey & Morgan, 1997), which could be an important factor behind the negative findings of aptitude by treatment interactions studies (e.g. Cronbach & Snow, 1977). In light of developments in theories of intelligence, and more strongly theory-linked batteries of intelligence, McGrew and Flanagan (1997) stress that specific cognitive abilities may indeed have important implications in student learning and achievement. As a result, scholars advocate for testing practices that include the full spectrum of cognitive abilities. Some research has begun to demonstrate the importance of more specific cognitive abilities in explaining academic difficulties in domains such as mathematics (Floyd, Evans, & McGrew, 2003), reading (Evans, Floyd, McGrew, & LeForgee, 2002; Flanagan, 2000) and writing (Floyd, McGrew & Evans, 2008).

Cross-Battery Assessment Knowledge Basis

Cross-battery assessment is based on three important areas of information, termed "pillars" (Flanagan & McGrew, 1997; Flanagan, Ortiz & Alfonso, 2007; McGrew & Flanagan, 1997). The foundations of cross-battery assessment are designed to maximize content validity by reducing construct irrelevant variance and construct underrepresentation (Flanagan, Ortiz & Alfonso, 2007). An assessment that is invalid due to construct underrepresentation is too narrow and does not sample from important dimensions of the construct it attempts to measure (Merrick, 1995). For intelligence tests, construct underrepresentation can occur when an assessment fails to measure important cognitive abilities (Flanagan & McGrew, 1997). Conversely, construct irrelevant variance is found in assessments that are too broad and inadvertently include constructs separate from the target variable(s) in their measurements (Merrick, 1995). Construct irrelevant variance can create interpretational difficulty by confounding multiple constructs within one measurement score. If a subtest measures more than one ability then it can be difficult to understand which ability most influenced an examinee's performance on that test.

Cross-Battery Pillar I. Cross-battery assessment's first pillar rests on the use of CHC theory as a guiding framework. Using CHC theory to guide assessment allows clinicians to construct assessments that more fully represent all human cognitive abilities. Moreover, since most current major test batteries are based on CHC theory, it represents a unifying framework that is applicable to nearly all modern intelligence batteries.

Cross-Battery Pillar II. Confirmatory factor analyses have allowed researchers to classify subtests into the broad abilities described by CHC theory. These classifications represent the second pillar of cross-battery assessment and a mechanism by which clinicians can protect against invalid assessment due to construct underrepresentation and construct irrelevant variance (Flanagan, Ortiz & Alfonso, 2007). Published references (e.g. Flanagan, McGrew & Ortiz, 2000; Flanagan, Ortiz, & Alfonso, 2007; McGrew & Flanagan, 1998) provide guidelines by which clinicians can select major subtests that measure broad abilities outlined by CHC theory. Flanagan et al. (2007) note that valid cross-battery assessments rely on these classifications in order to include strong, empirically validated measures of abilities in assessments. These authors suggest that subtests appearing to be mixed measures of two or more different abilities introduce construct irrelevant variance into ability scores, and therefore should be avoided for assessment purposes.

Cross-Battery Pillar III. Classification of subtests into narrow abilities represents cross-battery assessment's third pillar (Flanagan, Ortiz & Alfonso, 2007). Similar to pillar II, these classifications are important when guarding against construct underrepresentation in assessment, but at this level construct underrepresentation refers to the measurement of narrow abilities within broad ability factors. Because many intelligence tests include ability scores constructed from two or more subtests, it is critical to ensure that these subtests measure different narrow abilities to make certain that they adequately assess the broad ability they were intended to measure. For instance, inductive and deductive reasoning are part of the fluid intelligence construct. If a test battery includes two fluid intelligence measures, but both measure inductive reasoning, they may not adequately represent the fluid intelligence construct. Cross-battery assessment guides advocate for broad abilities to be assessed by two subtests measuring different narrow abilities. In contrast to broad ability classification, the classification of subtests into narrow abilities appears to be heavily based in evidence derived from expert consensus studies (e.g. McGrew, 1997; McGrew & Flanagan, 1998). These consensus studies frequently involve asking experts in CHC theory to categorize subtests by reading task descriptions and CHC ability definitions. Participants select which definition applies to a description, and researchers calculate interrater reliability using a percentage of agreement among participants.

Cross-Battery Assessment Guiding Principles

Effective cross-battery assessment requires clinicians to follow six guiding principles that allow for the gathering of psychometrically and theoretically sound results (Flanagan, Ortiz & Alfonso, 2007). First, clinicians should select an intelligence battery that addresses referral concerns. Examples of frequently used batteries include the Wechsler Intelligence Scale for Children – Fourth Edition (WISC-IV), the Woodcock-Johnson Psychoeducational Assessment Battery – Third Edition (WJ-III Cog), or the Stanford Binet – Fifth Edition (SB-V). Second, clinicians should use subtests from a single battery to represent CHC abilities whenever possible. This allows for results to be derived from actual norming data. Third, clinicians should use subtests that have been classified into broad and narrow ability categories through acceptable methods. Flanagan and colleagues (2007) suggest that factor analytic and expert consensus studies are appropriate types of data to use when selecting subtests. Use of empirical data when selecting subtests allows clinicians to select relatively pure ability measures that assess different narrow abilities. A minimum of two subtests are required to adequately assess a broad ability (Flanagan, Ortiz & Alfonso, 2007). Fourth, cross-battery assessment experts underscore that if a test battery does not include at least two indicators of a broad ability (that each measure different narrow abilities), then the battery should be supplemented by at least two indicators of a broad ability from another battery (Flanagan, Ortiz & Alfonso). Fifth, when a battery does have to be supplemented with other tests, it is important to select subtests from a battery that has been developed and normed in close proximity to the core battery. This principle is designed to guard against the "Flynn effect" or the finding that individuals tend to score higher on test batteries with older norms (Flanagan, Ortiz & Alfonso, 2007). Finally, the last principle of cross-battery assessment warns clinicians to use the smallest number of batteries necessary to minimize extraneous variance due to differences in characteristics of different norming samples.

Interpretation of Cross-Battery Assessments

Because cross-battery assessment often requires a reorganization of test constructs, it also requires the use of alternative procedures in order to interpret examinee performance from a CHC framework (e.g. Flanagan, Ortiz & Alfonso, 2007). CHC proponents stress that the interpretation of cross-battery assessment data begins with an assumption that examinee performance falls within the normal range of functioning (i.e. +/- one standard deviation from the normative mean) in order to avoid a predisposition to see areas of disability in test data. Such a stance is similar to the estabilishment of a null hypothesis in statistical procedures, and is important in avoiding a bias to see disability where one does not exist. Hypotheses are then created from an understanding of referral questions, using research as a means of hypothesis generation. For instance, some research suggests that broad abilities may demonstrate important effects on academic achievement (e.g. Evans, Floyd, McGrew, & LeForgee, 2002; Flanagan, 2000; Floyd, Evans & McGrew, 2003). Therefore, if an examinee is referred for reading difficulties, assessment should naturally focus on abilities that are empirically supported to be important to the reading process. Once abilities have been targeted for assessment, clinicians can begin to construct an appropriate test battery. After scoring all administered tests, the examiner can begin to make normative comparisons.

Cross-battery assessment guidebooks (e.g. Flanagan, Ortiz & Alfonso, 2007) typically emphasize the interpretation of broad, stratum II abilities. Broad abilities can be interpreted so long as they 1) are represented by at least two subtests that measure two different narrow abilities and 2) reflect a unitary construct (i.e. operationally determined by the lack of statistical significance between two or more subtest scores). If all broad abilities are interpretable, then it is possible to determine whether findings support assessment hypotheses. If a broad ability does not appear interpretable, then it may be necessary to gather additional data by administering a second subtest that measures the same narrow ability as the lower score in the construct (important when determining if the score is indeed a deficit). More data on an uninterpretable broad ability is necessary when the ability is central to the referral question and one of its narrow abilities falls below normal limits.

The Purpose of the Study

With an understanding of cross-battery interpretation procedures, it is possible to see that the classification of subtests into broad and narrow ability categories are both critical foundations of cross-battery assessment (i.e., Pillar II and III). However, the methods used to classify subtests into broad and narrow ability categories often vary. Both methods are associated with limitations. Broad ability classifications tend to be made using factor analysis. The use of factor analysis is limited when making narrow ability classifications because such a study would be extremely large, requiring a large sample size and numerous cognitive batteries. As an alternative, researchers have made narrow ability classifications by using expert consensus methods that rely on task analysis. Task analysis may confound observable demands such as task presentations and response requirements, with unobservable demands like examinees' use of cognitive abilities and mental processes. It may also be difficult to determine the most important cognitive ability measured by a subtest through task analysis (Frisby & Parkin, 2007).

Because “a proper multivariate study of both narrow and broad characteristics of all tests would be a daunting task” (Phelps, McGrew, Kopitk & Ford, 2005, p. 69), more investigation is needed to determine the viability of using task analysis to classify subtests into narrow abilities. The purpose of this study is to evaluate similarities and differences in the category outcomes that result from these two data types (task data and ability data) to ascertain whether task analytic methods can provide results similar to an ability analysis. It uses a group of six subtests previously classified as measuring narrow abilities subsumed by fluid intelligence (*Gf*) to compare similarities and contrast differences in the results of a an ability based analysis and a task analysis.

Currently, all major studies that attempt to classify subtests into broad abilities use a “battery-focused approach.” In this approach, a major intelligence battery is analyzed in terms of its fit with CHC factors, or joint analyzed with the Woodcock Johnson Test of Cognitive Abilities to provide markers for broad ability factors. To date, few studies have attempted an “ability-focused approach” where subtests from more than one or two batteries that are hypothesized to measure the same broad ability are subjected to confirmatory factor analytic methods. One reason may be that it would require the administration of more than two cognitive batteries, and such data is not readily available for analysis. Being much narrower in scope than past research, such a study may better validate narrow abilities measured by subtests from multiple test batteries.

The results of such a confirmatory factor analysis would serve as a standard from which to compare task analytic results of narrow ability classifications. If task analytic results appeared similar to ability-analysis results, it would add to the viability of using task analysis to make narrow ability classifications. To date, because most task analytic studies involve the use of measures of consensus, the present study uses an alternative methodology to increase the level of empiricism associated with task analysis. Cluster analysis is used to analyze task demands because its purpose is to make classifications based on similarities among observations (Hair & Black, 2000). Practitioners trained in intelligence testing provide data on the degree of similarity they perceive in the task demands of cognitive tests. This similarity data is subjected to cluster analytic methods and the cluster membership of subtests used in the factor analysis is tracked as clusters become more and more homogenous.

An understanding of whether these two data types lead to the same conclusions can provide practioners with more security in relying on the current knowledge base when interpreting broad and narrow cognitive ability measures. Similarly, it will allow researchers to better understand the relationship between the task demands of a cognitive task and the abilites it may measure.

CHAPTER TWO: REVIEW OF LITERATURE

History of Factor Analytic Theories of Intelligence and the Development of CHC Theory

Pioneered by Charles Spearman (1904), factor analysis is a statistical technique that partitions variance between variables (or mental tests in his research) into three different components: 1) common variance or variance shared with other variables, 2) specific variance, or variance that is not shared with other variables, and 3) error variance, or unreliable random variance (Bryant & Yarnold, 1995). As a result of his factor analytic studies, Spearman (1927) posited a two-factor theory of intelligence (Cohen & Swerdlik, 2002). Spearman's two-factor theory considered variance common to all tests as *"g"* or a general factor of intelligence. He defined *"g"* as a type of mental energy available for problem-solving (Cohen & Swerdlik, 2002; Sattler, 2001; Spearman, 1927). Variance unique to a test or cognitive task was considered to be a specific ability (as opposed to the general ability). Through factor analysis, variance that was unique to certain types of task, but not all tasks, was identified as "group factors." (Cohen & Swerdlik).

Spearman's work, and the identification of group factors, was a catalyst for the creation of "multifactor" theories of intelligence. Two multifactor theories, Cattell and Horn's Gf-Gc theory (Cattell, 1963; Horn, 1998; Sattler, 2001), and John Carroll's Three Stratum Theory of Abilities (Carroll, 1993) form the foundation of CHC theory (McGrew, 1997). Raymond B. Cattell, a student of Spearman's, began to describe his theory in the first half of the 20th century (Jensen, 1998). Gf-Gc theory originally described two factors: fluid intelligence and crystalized intelligence. Fluid intelligence refered to non-verbal, tests of mental power that did not depend on exposure to cultural information. Alternatively, crystalized intelligence represented acquired skills and knowledge that are dependent on exposure to cultural information (Sattler, 2001). Cattell and Horn later identfied other factors to describe the structure of intelligence after applying factor analysis to large datasets of diverse cognitive tasks (Horn, 1998, Jensen, 1998; Sattler, 2001).

It is important to note that Cattell and Horn's Gf-Gc theory does not outline a general ability factor (McGrew, 1997; Jensen, 1998). At the origin of the theory, when it only included two factors, a lack of a "g" factor was due to mathematical constraints (Jensen). However, even after the inclusion of other broad factors into the theory, Horn and Noll (1997) maintained that the structure of intelligence did not include a general factor. They stressed that different abilities displayed different rates of development and developmental declines, and also varied in their neurological functioning and level of heritability. Due to these different ability characteristics, Horn and Noll (1997) stressed that inclusion of a general ability factor did not adequately explain the structure of intelligence.

Carroll's (1993) *Three Stratum Theory of Cognitive Abilities*, a second basis for CHC theory, was derived from the reanalysis of over 460 data sets and ultimately presents evidence for three hierarchically-arranged strata of abilities. As a testament to the importance of Carroll's work, McGrew (1997) suggests that "all scholars, test developers, and users of intelligence tests need to become familiar with Carroll's treatise on the factors of human abilities" (p. 151). Carroll (1997) indicated that his theory bears resemblence to earlier multifactor theories of cognitive abilities because he used many of the same datasets in his analyses. He described his theory as a map of a large number of abilities, where their collective relationships can be derived by classifying them into strata of narrow, broad and general ability. The general factor, "g" represents the highest, third stratum in the hierarchy. Carroll's stratum II, broad factors included "abilities in the domain of," language, reasoning, memory and learning, visual perception, auditory reception, idea production, cognitive speed, knowledge and achievement, and micellaneous domains of ability, such as sensory and attentional abilities.

Gf-Gc Theory and Three-Stratum Theory form the basis of CHC theory (McGrew, 1997; Flanagan, Ortiz & Alfonso, 2007). There is much in common between these theories. Neverthelss, they vary in their interpretation of a number of abilities. McGrew (1997) reports differences between the two theories to include the existence of a third stratum or, *"g"* factor, and the classification of a number of traditional academic and memory abilities. He used subtests from the *Woodcock-Johnson Test of Cognitive Abilities – Revised* (WJ-R Cog; Woodcock & Johnson, 1989) to provide data for four models that he subjected to confirmatory factor analysis. McGrew designed model one to encompass Carroll's model, and it included six factors: 1) combined associative memory and memory span; 2) combined crystalized intelligence and reading/writing; 3) combined fluid reasoning and quantitative knowledge; 4) visual-spatial processing; 5) perceptual speeding and 6) phonetic coding, with a dual-loading on crystalized intelligence. His second model removed the dualing loading of the phonetic tests, and separated associative memory from memory span. McGrew's third model also separated reading and writing from crystalized intelligence. Finally, model four separated quantitative knowledge and fluid reasoning. Models 3 and 4 resulted in relatively equivalent fit statistics. However, McGrew noted that fluid intelligence and quantitative knowledge expressed distinct dvelopmental growth curves, which he interpreted as evidence indicative of separate constructs. As a result, he considered model 4 to be a more appropriate integration of the two theories.

The resultant CHC theory includes 10 broad abilities and over 70 narrow abilities. As outlined by Flanagan, Ortiz and Alfonso (2007), the broad abilities include:

Fluid intelligence (Gf) represents mental operations used in novel tasks that cannot be performed with any degree of automaticity. Novel tasks include stimuli or require methods of completion that examinees have generally lacked exposure to, or may never have performed before, where the fundaments are easy to understand by the widest range of cultural groups (e.g., pictures, abstract shapes and symbols). These tasks involve forming and/or recognizing concepts, identifying relationships, drawing out inferences, understanding implications and problem-solving;

Crystallized intelligence (Gc) refers to an individual's acquired knowledge of their dominant culture. These cognitive abilities are primarily verbal or language-based in nature;

Quantitative knowledge (Gq) represents acquired quantitative knowledge. Specifically, it involves the application and comprehension of mathematical concepts;

Reading/Writing ability (Grw) reflects a knowledge bank that includes reading and writing skills required for the expression of thought via writing and the comprehension of language through the written word;

Short-term memory (Gsm) requires the maintenance of information in immediate awareness and using it within a few seconds;

Visual processing (Gv) represents the ability to visualize, manipulate, analyze and synthesize visual patterns, stimuli or information.

Auditory processing (Ga) involves the perception of sound patterns, specifically the ability to analyze, synthesize and manipulate auditory stimuli, and especially discriminating subtle sounds, such as when under distraction;

Long-term storage and retrieval (Glr) represents the cognitive ability to store information in long-term memory and to retrieve it at a later time through association;

Processing speed (Gs) involves the rapid, fluid performance of simple, clerical-type tasks, especially when under pressure to maintain attention and concentration;

Decision/Reaction time/Correct decision speed (Gt) involves the ability to make immediate decisions or react to quickly changing stimuli.

Major efforts in cross-battery assessment literature have attempted to apply CHC theory to the classification of subtests from major test batteries. Thus, critical literature associated with cross-battery assessment revolves around studies classifying subtests from test batteries into broad and narrow ability factors, as outlined by CHC theory. These studies include single battery and joint battery factor analyses and expert consensus studies.

Applying CHC Theory to Major Cognitive Batteries Through Factor Analysis

CHC theory has been a major influence in modern day test development, starting with the *Woodcock-Johnson III* (WJ-III Cog; McGrew & Woodcock, 2001). The WJ-III Cog standardization sample provides the largest database of support for CHC theory from one source (McGrew & Woodcock, 2001). Other modern tests, including the *Stanford Binet Intelligence Scales, Fifth Edition* (SB-IV; Roid, 2003), the *Differential Abilities*

Scale, Second Edition (DAS-II; Elliot, 2007) and the *Kaufman Assessment Battery for Children, Second Edition* (KABC-II; Kaufman & Kaufman, 2004), have also been based on the CHC ability taxonomy.

Much of the research validating the use of the CHC taxonomy for the interpretation of intelligence scores come from both single battery confirmatory factor analyses and joint confirmatory analyses including more than one intelligence battery (Flanagan, McGrew & Ortiz, 2000). Single battery confirmatory factor analyses are often conducted by independent researchers as a way of assessing whether alternative factor structures fit a battery better than the one indicated by a test's authors. They may also be included in test manuals as a way of confirming content validity (e.g., Hammill, 1998; Roid, 2003). Though single battery analyses are important, often the number of subtests in single battery analyses proves insufficient to allow all the abilities tapped by subtests to emerge during analysis (Woodcock, 1990; Flanagan & McGrew, 1998). Including more than one battery in a factor analysis provides more factor indicators, and allows subtests to load on constructs defined theoretically. Thus, as an alternative method for investigating a test battery's factor structure and identifying the abilities measured by subtests, researchers have conducted confirmatory factor analyses on a number of major test batteries by using WJ-III Cog subtests as marker variables for CHC Stratum II abilities (Flanagan & McGrew, 1998; Keith, Kranzler & Flangan, 2001; Phelps, McGrew, Knopik & Ford, 2005; Sanders et. al., 2007; Tusing & Ford, 2004; Woodcock, 1990). Both methods demonstrate the importance of CHC theory when interpreting test data, and also highlight tests that measure fluid intelligence as a Stratum II ability.

Single Battery Confirmatory Factor Analyses. Recent independent confirmatory factor analyses have demonstrated that CHC theory may actually provide a better, more interpretable structure for scores from a number of IQ test batteries (Keith et. al., 2006; Kranzler & Keith, 1999). Validity studies in the WISC-IV manual (Wechsler, 2003) suggest that WISC-IV subtests measure *"g"* and subsequently, Verbal Comprehension (VCI), Perceptual Reasoning (PRI), Working Memory (WMI) and Processing Speed (PSI) factors. However, the manual's four factor structure has been contested by independent analyses endorsing a CHC hierarchy that requires modification to traditional WISC-IV interpretation (Keith et al., 2006). For instance, the PRI factor may not actually reflect a unitary construct. Instead it appears to consist of a mix of both fluid intelligence (*Gf*) and visual processing (*Gv*) abilities. Block Design and Picture Completion subtests appear to load on visual processing, while Matrix Reasoning and Picture Concept subtests load on fluid reasoning. Keith and colleagues demonstrated that a model including the WISC-IV PRI factor reflects a poorer fit to the standardization data compared to a model with a separte fluid reasoning and visual-spatial processing factor (Keith et al., 2006, p. 117). Additionally, the researchers suggest that the Arithmetic subtest appears to be a primary measure of fluid intelligence, while also acknowledging that the subtest likely measures a complex set of abilities (p. 118).

Kranzler and Keith's (1999) confirmatory factor analysis suggests that the Cognitive Assessment System (CAS) may contain similar mixed constructs. Based on Planning, Attention, Simultaneous, and Sequential (PASS) theory (Naglieri & Das, 1997), the CAS stresses the importance of three types of cognitive abilities or systems when processing information: a) planning involves the formation, selection and monitoring of strategies and plans of action; b) attention represents the ability to allocate cognitive resources and effort to tasks; c) information processing, comprised of simultaneous and successive processing abilities, encompasses the acquisition, storage and retrieval of information from the environment; lastly. While initial analyses reported in the CAS manual (Naglieri & Das, 1997) seem to suggest that the PASS structure describes CAS subtests well, Kranzler and Keith's analysis suggests that a hierarchical model may be a better fit. Similar to the WISC-IV's PRI construct, the CAS simultaneous processing ability (i.e. Nonverbal Matrices, Verbal Spatial Relations and Figural Memory) may actually represent a complex mix of *Gf* and *Gv* CHC Factors (Kranzler & Keith, 1999).

Other intelligence batteries include confirmatory factor analyses to confirm CHC theory as part of their validity evidence, including the Stanford-Binet, Fifth Edition (SB-V; Roid, 2003) and the Woodcock–Johnson III Tests of Cognitive Abilities (Woodcock, McGrew, & Mather, 2001). The SB-V measures five of the ten CHC broad abilities: fluid intelligence, quantitative reasoning, crystallized knowledge, short-term memory and visual processing. This five-factor model was judged to be the best fit during a series of confirmatory factor analyses. The authors of the WJ-III cognitive battery also endorsed a CHC model for their test (Woodcock, McGrew & Mather, 2001). Their data suggest that the WJ-III Cog measures *Gc, Gq, Glr, Gv, Ga, Gf, Gs* and *Gsm* abilities. It is important to note that most of the subtests included in the WJ-III cognitive battery appear to be relatively pure measures of their respective abilities, and thus, the WJ-III Cog appears to be free from extraneous, construct irrelevant variance. Additionally, it appears that this factor structure is supported across age groups in the norm sample (Taub & McGrew, 2004).

Joint-Battery Confirmatory Factory Analysis. Due to a limited number of subtests contained in test batteries, single battery confirmatory studies may not be able to ascertain the usefulness of CHC theory for the interpretation of a single battery. McGrew, Flanagan and Ortiz (2000) provide an example of the limitations of single battery factor analysis based on Woodcock's (1990) discussion of the WJ-R and the WISC-R. Woodcock's conclusions not only demonstrate the usefulness of CHC theory, but stress the critical importance of content validity in the construction of tests. Factor analysis of the WISC-R suggests that it contains Verbal Comprehension (VC), Perceptual Organization (PO) and Freedom-From-Distractibility (FFD) factors. While VC and PO appear to measure *Gc* and *Gv* abilities respectively, FFD does not appear to be associated with any type of *Gf-Gc* ability. Joint factor analysis of the WISC-R with the WJ-R demonstrates that the FFD factor does not emerge when its subtests are allowed to load on factors with subtests from the WJ-R. Because inclusion of WJ-R subtests provides more indicators for the detection of abilities measured in the WISC-R, rather than loading together on one factor, FFD subtests load on factors that appear to represent short-term memory (*Gsm*), quantitative knowledge (*Gq*) and processing speed (*Gs*). As a result, some scholars conclude that the FFD factor likely represents "the fallout from the weak substantive or theoretical foundation of the Wechsler Scales" (Flanagan, McGrew & Ortiz, 2000, p. 72).

Joint factor analytic studies typically use the Woodcock Johnson series of cognitive tests co-analyzed alongside a target battery. Researchers have reported analyses of Woodcock subtests with the *Stanford-Binet, Fourth Edition* (Thorndike, Hagen & Sattler, 1986), the *Wechsler Adult Intelligence Scale*, the *Wechsler Intelligence Scale for Children – Revised (1974)* and the *Kaufman-Assessment Battery for Children* (Kaufman, 1983; Woodcock, 1990), the *Detriot Test of Learning Aptitude, Third Edition (*DTLA-3; McGhee, 1993), the *Detriot Test of Learning Aptitude – Adult (*DTLA-A; Hammell & Bryant, 1991; Buckholt, McGhee & Ehrler, 2001), the *Kaufman Adult and Adolescent Intelligence Test (*KAIT; Kaufman & Kaufman, 1993; McGrew & Flanagan, 1998), the *Cognitive Assessment System* (CAS; Naglieri, & Das, 1997; Keith, Kranzler & Flanagan, 2001) and the *Differential Ability Scales* (DAS; Elliot, 1990; Sanders, McIntosh, Dunham, Rothlisberg & Finch, 2007). All studies indicate that CHC theory can be used in the interpretation of these intelligence tests. Also endorsed by McGrew (1997), Woodcock (1990) used a rating system to classify subtests as strong, moderate or mixed measures of broad abilities. Strong subtests demonstrated a factor loading greater than .500 and a secondary factor loading that was less than one-half of the primary loading. Moderate subtests were classified with two different criteria: a) subtests loading on a primary factor under .500 and also loaded on a second factor with a loading that was less than one-half of the primary loading, or b) subtests loading with any value on a primary factor that also measured a secondary factor with a loading between one-half and seven-tenths of the primary loading. Woodcock labeled subtests with a secondary loading greater than seven-tenths of a primary loading as “mixed,” as it can be difficult to distinguish a primary factor loading from these subtests.

Research across batteries consistently demonstrates that Woodcock-Johnson subtests *Analysis-Synthesis* and *Concept Formation* appear to be strong measures of fluid intelligence (i.e., *Gf*) and underscores their usefulness as markers for the *Gf* broad ability. In Woodcock's analyses of the WJ-R (1990), these subtests demonstrated loadings of .586 and .682 on a *Gf* factor. Similarly, as part of the WJ-III Cog, *Analysis-Synthesis* and *Concept Formation* demonstrated loadings of .72 and .73 when analyzed with the DAS (Sanders, McIntosh, Dunham, Rothlisberg & Finch, 2007) and comparable loadings when analyzed with the CAS (Keith, Kranzler & Flanagan, 2001). Additionally, both subtests loaded in the .60's when analyzed with subtests from the WISC-III (Phelps, McGrew, Knopik & Ford, 2005).

Due to *Analysis-Synthesis'* and *Concept Formation's* strong *Gf* loadings, subtests that load with them in joint analyses are also important to note as measures of fluid intelligence. Woodcock (1990) demonstrated that the SB-IV's *Matrices* subtest loaded with these subtests (.609) and validity data for the Stanford-Binet, Fifth Edition (Roid, 2003) indicates that its *Nonverbal Fluid Reasoning* and *Verbal Fluid Reasoning* subtests demonstrated strong loadings with *Analysis-Synthesis*. KAIT subtests *Logical Steps* and *Mystery Codes* demonstrated factor loadings in the .90s with these subtests (Flanagan & McGrew, 1998) and Keith and colleagues (2001) CAS/WJ-III Cog analysis indicated that the WJ-III's *Numerical Reasoning* subtest loads highly (.74) on a *Gf* factor. Phelps et. al (2005), who included WJ-III supplemental tests in their analyses, demonstrated that *Number Matrices*, *Number Series* and *Applied Problems* can be considered tests of *Gf* as well. Sanders et al. (2007) demonstrated that DAS subtests *Matrices* and *Sequential and Quantitative Reasoning* measure aspects of fluid intelligence, both demonstrating loadings in the .70s with *Concept Formation/Analysis-Synthesis*.

Other subtests appear to be more mixed measures of fluid intelligence. Woodcock's (1990) analyses with the WJ and WJ-R demonstrated that *Spatial Relations* appears to be a moderate measure. *Verbal Analogies* appears to be a mixed measure with a moderate loading on *Gf* and *Gc*. The DTLA subtest *Symbolic Relations* appears to load on *Gf, Gq* and *Ga* (Buckhalt, McGhee & Ehrler, 2001). Phelps et. al (2005) demonstrated that the WJ-III *Planning* subtest also appears to be a moderate measure of fluid intelligence.

Narrow Ability Measures

Because narrow abilities represent the building blocks of composite scores under the cross-battery assessment model, it is critical to understand which narrow abilities subtests may measure. While factor analytic studies like those mentioned earlier can identify a battery's structure and highlight subtests that measure different broad abilities, they are unable to identify more specific abilities assessed by subtests with few exceptions. This inability is primarily due to difficulty administering enough subtests to a large enough sample of examinees in order to have multiple indicators of a narrow ability. Current studies would have to greatly increase both the number of tests adminstered to participants and also the number of participants included in their sample size.

Identifying Narrow Abilities Through Factor Analysis. A small number of the reviewed joint factor analysis studies contained enough ability indicators to provide evidence narrow ability classifications for some fluid intelligence subtests' narrow ability classifications. Flanagan and McGrew (1998) demonstrated that the fluid intelligence factor that emerges in WJ-III Cog/KAIT joint analysis may actually represent two narrow abilities, general sequential reasoning (including *Analysis-Synthesis* and *Logical Steps*) and induction (*Concept Formation* and *Mystery Codes*). Phelps and colleagues (2005) also demonstrated WJ-III Cog and WISC-III fluid intelligence subtests may represent narrow ability measures as well. After creating a place holder factor that included WJ-III Cog's *Analysis-Synthesis* and *Concept Formation* subtests, a quantitative reasoning narrow ability emerged, loading on fluid intelligence and including *Planning, Number Matrices, Number Series* and *Applied Problems* subtests from the WJ-III Supplemental and Achievement tests. McGrew and Woodcock (2001) present a three-stratum model for the WJ-III battery which includes both cognitive and achievement tests. They highlight a number of narrow abilities in the model, including the *Gf* narrow ability, Quantitiatve Reasoning (RQ). Their analyses suggest that subtests such as *Analysis-Synthesis, Applied Problems* and *Quantitaitve Concepts* may load on an RQ factor.

Task Analytic Methods in Subtest Classification

Because of the high number of participants and ability indicators required to identify narrow abilities through factor analysis, researchers have used other methods to understand the abilities tapped by cognitive subtests. Task analysis involves the comparison of subtests' functional task demands: their mode of stimuli input and response output, as well as any required strategies and cognitive processes that are necessary for successful completion of their test items. Clinicians may use task analysis to understand examinee cognitive strengths and weaknesses, while researchers may use it to label factors common among a number of subtests.

Spearman engaged in task analysis to provide context to the results of factor analytic studies by considering the similarities between what he considered high g-loaded, versus low g-loaded tests (Jensen, 1998). From his comparisons, Spearman outlined three "laws of noegenesis" or the "production of new knowledge, or mental content from sensory or cognitive experience" (Jensen, 1998, p. 35). As reported by Jensen, Spearman labeled these laws apprehension of experience, eduction of relations and eduction of correlates. To Spearman, apprehension of experience refered to an awareness of perceiving the characteristics of stimuli. Eduction of relations expressed the tendency for the presentation of two or more stimuli (called "fundaments") to "draw out" an appropriate comparison between them. For instance, upon presention of the word stimuli "red-blue", the examinee would educe the relationship "color." Spearman's eduction of correlates expressed that the presentation of a fundament and a relationship together would result in the tendency to make correlates of that relationship salient. Presentation of the fundament "small" along with the relationship "opposite" would prompt the examinee to educe a correlate of these concepts and result in the response "big." Spearman suggested that tasks with a high *"g"*-loading would be tasks that require the eduction of relations and the eduction of correlates.

Though Spearman described some characteristics of items that may better measure "g" through the task analysis process, Jensen (1998) stressed that "g" cannot be described by the specific functional requirements of a test. Results from diverse types of cognitive tests generally "rank order persons in much the say way, despite the tests' often vastly different appearance in information content and form of response" (Jensen, 1992, p. 175). This tendency defines Spearman's theorem of the indifference of the indicator (Jensen, 1992; 1998), a principle evident within modern test batteries. For instance, within the WISC-IV, *Block Design*, a measure of examinees' ability to replicate puzzles with colored blocks, and *Vocabulary* a measure of the ability to supply definitions to words, correlate highly with the FSIQ (.70 and .79 respectively). However, their correlation with each other is .48 (Wechsler, 2003b). These tests have very different task demands, and a task analysis may suggest that they measure different constructs, nevertheless, each test correlates very highly with a general factor.

Though the indifference of the indicator emphasizes a common construct measured across cognitive tasks, group factors can be described in terms of "the obvious characteristics of the kinds of tests that load on them (such as verbal, numerical, spatial visualization, memory, mechanical…)" (Jensen, 1998 p. 91). This process is used to define and label factors based on the factor loadings of ability tests. As Jensen has suggested, it is important when interpreting factor analytic results of broad abilities. Thus, though the concept of "*g"* does not highlight the importance of task analysis, task analysis is critcal when understanding group factors, or CHC broad ability factors.

Clinicians frequently engage in task analysis to understand examinee performance on cognitive subtests and to develop hypotheses of cognitive strengths and weaknesses. To facilitate this process, test developers often publish test supplements that discuss potential reasons for differences in performance. Numerous books and supplementals are written to aid clinicians in how to interpret examinee scores, analyze examinee performance and use test batteries appropriately with diverse types of populations (e.g., see Flanagan & Kaufman, 2004; Schrank & Flanagan, 2003). Schrank and Flanagan (2003) stress the importance of understanding task demands when understanding subtest scores:

"Examiners should pay particular attention to the description of task demands across test or subtest items. It is important to consider the multidimensional nature of

> *the task requirements. For example, a cancellation task...involves or requires several abilities and processes including sustained attention, processing, motoric speed, and executive functioning,. . . . Consider the task demands of this test [a cancellation task]. The examinee needs to scan a page of items heavily laden with little pictures (soccer balls, dogs, tea cups) and is required to locate and mark a repeating pattern (ball followed by a dog) in rows of pictures that are purposely designed to be visually 'busy.' This test has a 3 minute time limit." (p. 51, 68).*

Analyzing task demands allows practitioners to regroup subtests in attempt to discover areas in which examinees may show strengths and/or deficits. Interpretations of subtests can appear "fairly similar, but [diverge] in accordance with each author's personal orientation" (Kaufman, 1994, p. 49). Different theories may suggest that subtests are similar or different from each other based on the task demands they impose on the examinee. As Kaufman (1994) demonstrates, subtests may be grouped based upon their sensitivity to attention and concentration, motor demands, visual organization, lack of motor coordination, or any other kind of functional requirement. In all likelihood the number of variations and combinations of subtest task demands has no limit.

History of Task Analysis in Test Interpretation

Frisby and Parkin (2007) argue that task analysis represents a rational, but subjective process. During the early years of test development and interpretation, subtests were selected for inclusion into test batteries through an "armchair" intuitive analysis of similarities and differences in their functional demands. Individually, examiners considered subtest characteristics and selectively grouped together tests that appeared similar. For instance, Wechsler originally organized his test battery dichotomously "based on intuitive and rational considerations" (Kaufman, 1979, p. 130). Subtests requiring verbal responses from examinees were considered to be verbal in nature, while those requiring pointing or other types of motor responses were considered more performance oriented (Kaufman, 1979; Kaufman, 1994; Kaufman & Lichtenberger, 2002). However, as Kaufman notes, Wechsler expected that "the abilities represented in the tests may also be meaningfully classified in other ways" (Wechsler, 1974, p. 9). As he predicted, researchers and scholars have outlined numerous ways to organize cognitive subtests (Kaufman & Lichtenberger, 2002). Similar to Wechsler's original methods, most of these organizational schemes are rational analyses of subtests' task demands.

Reorganization of the Wechsler Scales. Throughout their history, the Wechsler scales have been reorganized by numerous scholars and researchers, including Rapaport and colleagues (Rapaport, Gill & Schafer, 1945), Bannatyne (1971), and Kaufman (1979). Rapaport and colleagues divided the Wechsler-Bellevue (W-B) scale into verbal and performance halves, and further subdivided those categories based on their clinical experience administering those tests (Kaufman). Within the verbal category, they stressed that *Comprehension, Information* and *Similarities* required significant verbal ability, while *Arithmetic* and *Digit Span* tapped attention and concentration abilities. They divided the performance category into tasks that were more (*Block Design, Object Assembly, Digit-Symbol Coding)*, or less dependent on motor skills.

Bannatyne's (1979) reorganization of the original WISC had both similarities and differences compared to Rapaport's categorization. Similar to Rapaport, Bannatyne outlined a verbal ability category (*Similarities, Vocabulary* and *Comprehension*), though his method of categorization differed from Rapaport regarding other subtests. Bannatyne stressed that *Information, Vocabulary* and *Arithmetic* reflected an acquired knowledge category. *Picture Completion, Block Design* and *Object Assembly* were considered measures of spatial ability, while *Arithmetic, Digit Span*, and *Coding* subtests reflected sequencing skills.

Kaufman (1979) provides a detailed discussion of ‘profile attack’ via rational analysis by building, and expanding on these clinicians’ numerous methods of categorization. Kaufman (1979, Table 3-1) suggested a number of recategorizations of WISC-R subtests, some based on a multifactor theory of intelligence outlined by Guilford, and others derived from rational interpretation. He asserted that the Performance Scale consisted almost exclusively of Guilford's evaluation operation with figural content. Alternatively, WISC-R verbal subtests were mainly comprised of semantic content, with only the *Comprehension* subtest requiring evaluative processes from the examinee. Subtests that make up the Freedom from Distractibility Index (*Arithmetic*, *Digit Span* and *Coding*) all utilized symbolic content.

Kaufman (1979) outlined a number of additional intuitive or theoretical subtest categorizing methods for the WISC-R. They are not associated with any of the particular classification schemes previously discussed, but nonetheless demonstrate alternative ways to consider patterns of high and low scores on intelligence test profiles.

Patterns on the Verbal Scale included subtests requiring reasoning versus recall processes, subtests providing brief stimuli or long stimuli for the examinee, and subtests that require brief expressions in response from an examinee, or those that require lengthy verbal expressions.

Reasoning vs. Recall Processes. Similarities, Arithmetic, and *Comprehension* are verbal subtests that require reasoning, specifically problem-solving or the application of old, previously learned knowledge to novel situations. Alternatively, *Information,* *Vocabulary,* and *Digit Span* reflect subtests that merely require the retrieval of information stored in memory. However, within these categories, subtests are still not necessarily homogenous in their task demands. For instance, within the reasoning category, *Similarities* and *Comprehension* require verbal reasoning from examinees. *Arithmetic*, as the third subtest in the category, requires reasoning with numbers. To complicate this category further, *Arithmetic* and *Comprehension* require reasoning in social situations. *Arithmetic* uses mathematical word problems applied within real-life scenarios, while *Comprehension* requires reasoning around social norms. To some clinicians, *Similarities* may not appear as social in its task demands because the nature of its responses are more abstract and involve a relationship between words or concepts. Subtests in the recall category are also not homogenous in their task demands. Specifically, *Information* and *Vocabulary* require examinees to use long-term memory stores, while performance on *Digit Span* is more dependent on short-term or working memory.

Length of Stimuli/Length of Response. Verbal subtests can also be partitioned into categories based on the length of their stimuli, and the length of the response they require from the examinee (Kaufman, 1979). *Information, Arithmetic,* and *Comprehension* use lengthy verbal statements as part of their task demands. In contrast, considering that *Similarities* presents two words to the examinee, *Vocabulary*, one word, and *Digit Span*, a progressively longer series of numbers, their stimuli are relatively shorter. When categorizing these subtests based on the length of response required from the examinee, *Similarities, Vocabulary*, and *Comprehension* are considered similar because all require lengthy expressions from the examinee. Responses tend to be rationalizations, or definitions. Comparatively, *Information, Arithmetic*, and *Digit Span* require short answers such as a number, a specific response like a geographical location or historical figure, or the repetition of a string of numbers.

Kaufman (1979) also delineates classification schemes for subtests on the WISC-R's Performance scale. Bannatyne's Spatial Ability is one of the category types Kaufman mentions, although cognitive styles, processing and stimuli type, and imitation/problem solving dichotomies are also subtest task demands that can divide Wechsler Performance subtests. The WISC-R Performance scale also includes subtests that differ based on their problem solving requirements, the nature of their stimuli, and the mode of their response requirements (Kaufman).

Cognitive Style. Regarding cognitive styles, Kaufman suggests that the Wechsler's verbal scale consists of language tasks dominated by the brain's left hemisphere. Subtests on the Performance scale can be split between right brain functioning, and those requiring an integration of tasks for the two hemispheres. Kaufman argues that *Picture Completion* and *Object Assembly* require holistic visualization for successful completion. Other subtests on the scale (ie. *Picture Arrangement, Block Design, Coding* and *Mazes*) require more than just visual-spatial skills; they require considerable verbal instructions by the examiner, sequencing, or other analytic operations usually performed by the left side of the brain.

Processing and Stimuli Type. When analyzing task demands, Kaufman (1979) also makes distinctions between types of cognitive processing. Successive processing features serial and sequential interpretation of test stimuli, meaning that task completion requires the ability to understand that parts of a task must occur in an order. *Picture*

Arrangement, Coding, and *Mazes* represent subtests on the Performance scale that require examinees to sequentially engage with one part of stimuli at a time, be it the order of pictures, successive corridors in a maze, or a progression through a series of abstract shapes. Simultaneous processing assumes that all stimuli in a test are being considered at the same time. Accordingly, on the WISC Picture *Completion, Block Design*, and *Object Assembly* can be considered as subtests requiring simultaneous processing as part of their task demands. Kaufman (1979) notes that the simultaneous group is identical to Bannatyne's Spatial category and is very similar to the subtests considered to be right brain oriented.

Response Requirements. Block Design and *Coding* require examinees to imitate and/or reproduce test stimuli as they complete items on the subtest. Response requirements may also include tests that requiring pointing versus providing a verbal response.

Problem-solving Style. Performance subtests demand a range of problem solving sophistication from examinees. Problem solving may range from deducing the missing part of a picture to sequencing a group of story boards to form a coherent story. Some Performance tests consider of meaningful or concrete stimuli. *Picture Completion, Picture Arrangement,* and *Object Assembly* are subtests with identifiable (meaningful) stimuli that examinees likely have encountered in real life and are familiar with. In contrast *Block Design* and *Coding* stimuli are more abstract in nature. Note that this dichotomy can apply to certain subtests on the Verbal scale. Kaufman (1979) suggests that if reasoning with abstract stimuli is a strength for an examinee, clinicians may observe high *Arithmetic* and *Digit Span* scores, because numbers are also symbolic in nature.

Examinee Response Mode. Lastly, examinee response modes reflect another way to organize subtests on the Performance scale, similar to the scheme endorsed by Rapaport and his colleagues (Rapaport, Gill, & Schafer, 1945). *Block Design, Object Assembly, Coding,* and *Mazes* reflect subtests that require visual-motor responses, while *Picture Completion* and *Picture Arrangement* reflect responses comprised of specific visual organizational patterns.

Limitations with Rational Classifications. The multiple classifications derived from these same set of tasks reflects an important limitation associated with the rational analysis of subtests. Primarily, different aspects of a task may be more salient to clinicians depending on theoretical orientation. For instance, clinicians have categorized *Vocabulary* to reflect verbal comprehension ability because it requires examinees to respond to short, verbal stimuli with long verbal responsese. Clinicians have also categorized the subtest to reflect an acquired knowledge store because successful performance on the task requires access to long-term storage to retrieve word definitions. Though both abilities may be important in the completion of the task, it is difficult to know which ability may be the strongest influence on an examinee's task performance.

Task Analysis and Information Processing Theories

While early uses of task analysis represented efforts to increase the utility of subtest interpretation, more modern use of task analysis reflect an effort to apply models of information processing to cognitve assessment. Information processing reflects the sequences of mental operations and their outcomes that occur while an individual is engaged in a cognitive task (Sternberg, 1981). Information procesing models often include components such as a sensory register, a temporary holding space for stimuli gathered from sense organs, short-term memory, limited capacity stores for information in immediate awareness, and long-term memory, relatively permanent storage of knowledge (Floyd, 2005). Floyd stressed that an integration of information processing models with factor analytic research on cognitive abilities is necessary to a) increase collaboration between cognitive psychology and psychometric researchers, b) improve methods to gather evidence of validity for cognitive tests, c) facilitate the diagnosis of learning disabilities, and d) explain reasons for individual variation on cognitive tasks with greater precision. Task analysis represents a critical vehicle for this integration because it involves analyzing and describing characteristics of cognitive tasks.

Atkinson and Shiffirin (1968) presented an initial model of information processing that is now considered the "granddaddy" of information processing research (Floyd, 2005). Their model describes information or stimuli from the environment passing into awareness through a sensory register where it is stored only for a brief period of time. If the information is not attended to it is lost. Otherwise the information enters modality-specific short-term memory (e.g. auditory-verbal-linguistic, visual and touch-related memory stores). The use of storage strategies (such as reherasal) may transfer information from short-term to long-term memory or a store of acquired knowledge. This type of model inspired psychometric researchers to consider cognitive subtests from the perspective of the tasks they require of examinees and not exclusively by the abilities they may measure.

Carroll's Task Coding Scheme

While Kaufman used theWechsler series of tests when describing constructs used to categorize tests during task analysis, Carroll (1976) created a categorization scheme to define subtests as cognitive tasks in order to integrate theories of cognitive processing with factor analytic studies. He selected a sample of 74 cognitive tests included in the *Kit of Reference Tests for Cognitive Factors* (1963) that measured 24 different cognitive abilities. In order to understand these subtests as tasks rather than as a measure of some type of ability, Carroll tried to ignore, or "lay aside, and be unbiased by, any knowledge [he] had of the empirically determined 'factor structure' of each test" (Carroll, 1976 p. 37).

Carroll (1976; Floyd, 2005) outlined six characteristics useful in categorizing subtests as cognitive tasks. He specified that subtests may differ based on 1) the types of stimuli presented at task outset, 2) the overt responses examinees must make at the end of the task, 3) the structure of the task, 4) operations and strategies used to complete the task, 5) temporal aspects (such as duration) of any required operations or strategies an examinee must use to complete the task, and 6) the memory storage involved in the task (see Carroll, 1976, Table 1, p. 38). Each of these six broad classification schemes also has a number of criteria and sub-criteria.

Stimulus Materials. Subtests may consist of different types of stimuli, such as tangible items like blocks, words, pictures, or verbal phrases. Materials can also vary in the number of stimuli classes they use, their "completeness", and their interpretability. Some types of subtests may consist solely of one type of stimuli, while others may have more types. The WISC-IV's Vocabulary test allows the examinee to read a word and hear it pronounced by the examiner, thus providing two classes of stimuli for input: visual and auditory. By “completeness” Carroll accounts for subtests that may use planned interference as part of their task demands (i.e. progressively louder white noise to examinees to increase task difficulty in attending to target stimuli). For example, the *Auditory Attention* test from the WJ-III Cog , a task that requires the examinee to point at a picture based on a word that they discriminate from progressively louder white-noise, represents an example of this stimulus characteristic. Interpretability expresses that some subtest stimuli are unambiguous and immediately interpretable while other stimuli may be more ambiguous. Ambiguous in this sense does not consider stimuli as necessarily difficult to categorize and understand, more accurately it refers to stimuli that could be coded into examinees cognitive processes in several different ways. Other subtest stimuli may be anomalous and not immediately understood by examinees.

Examinee Responses. Carroll considered examinees’ responses to test items from several perspectives. They differ in the number and type of responses to be made, the mode of response, and the criterion that decides whether responses are deemed to be acceptable. Subtests may require examinees to select one response from a list of choices, create their own response based on stimuli and required mental operations, produce as many responses as possible, or respond a specific number of times. The method with which an examinee must produce a response may be through indicating a choice, producing a symbol or letter, writing a word, a phrase, a sentence, a paragraph (or even more), making a verbal response or creating a line or other type of drawing. When considering what criteria subtests use to determine the acceptability of responses, Carroll argued that answers could reflect numerous types of acceptability such as judgments or associations.

*Test Structure***.** Carroll used the term "task structure" in referring to time-related aspects of subtests (Floyd, 2005). For instance, many subtests may incorporate a time-delay into their task structure, where examinees are exposed to stimuli at one time, and then required to use the stimuli at a later time. In comparison, other subtests have a more "unitary" task structure, as they are not revisited at a later time.

Test Operations and Strategies. Operations and strategies refer to elementary information processes required of examinees when they are engaged in a task (Floyd, 2005). The operations and/or strategies that examinees may use to complete subtests may differ based on the number of operations or strategies that are required to complete the task, the type of operation needed, whether the operation is explicitly specified during task instructions, and how critical a particular operation or strategy is for successful test completion. Operations may also vary in the duration required to complete any particular cognitive task. Carroll (1976, see Table 1, p. 38) delineated 20 types of operations that examinees may use during subtest completion. These may include identifying similarities betweenstimuli, rotating a spatial configuration, storing items inmemory, constructing a hypothesis, or engaging in a specific visual search strategy.

*Temporal Aspects of Operation Strategie***s.** Operations employed during performance on a task may vary in duration based on task or individual examinee differences. They may also terminate themselves, such as when they lead to a correct solution, or they may not be self-terminating, as when time constraints require examinees to stop working on a particular test item.

Memory Storage. Lastly, Carroll (1976) differentiated memory store requirements of cognitive tasks. These requirements differ along three criteria: term, contents, and relevance of individual differences. “Term” refers to the length of memory and implies a specific memory type. Some tasks, such as reaction time tests, may only require stimuli to enter the sensory-register. Other tasks may require short-term memory. Carroll considered subtests that required examinees to remember information for a number of minutes to use intermediate-term memory. Other subtests may tap long-term memory. Carroll considered individual differences in his scheme, because some individuals, especially those with certain disabilities or neurological problems, may be severely impaired in their memory. Finally, Carroll (1976, Table 1, p. 38) listed 15 different types of content that can describe subtest stimuli.

Identifying Narrow Abilities Through Task Analysis.

Many narrow ability classifications appear to be based on expert consensus studies that rely on a task analysis (e.g. McGrew, 1997). McGrew asked ten scholars, all of whom were experienced with the development and interpretation of intelligence tests, to classify subtests into narrow ability categories. The primary ability presumed to be measured by a subtest was assigned a “1” rating by participants. If a participant felt that a subtest measured two primary abilities, they rated each ability with a “1.” If participants felt a subtest measured a lesser second ability, they rated the lesser ability with a “2”. Two or three experts completed these ratings for a number of major test batteries. As Glutting, Watkins and Youngstrom (2003) and McGrew (1997) report, these ratings lacked any type of interrater reliability measurement and have been modified numerous times (e.g. Flanagan, McGrew & Ortiz, 2000; McGrew, 1997; McGrew &

Flanagan, 1998). As a result, Glutting and colleagues stress that "placement of subtests within cross-battery factors was more a matter of speculative deduction than of demonstrable fact" (Glutting, Watkins & Youngstrom, 2003, p. 363).

Flanagan and colleagues (2001) used an expert consensus process to classify achievement-oriented subtests into CHC factors. They sent a group of CHC experts a list of broad and narrow ability definitions, and included an example of a task that could be classified within each narrow ability definition. Additionally, each expert received a packet containing approximately 40 test descriptions from a pool of 323 different tests. Without knowing the name of the test, or the battery it represents, experts selected one broad ability and one narrow ability that they felt was best reflected by the test description.

To evaluate the results of the expert classifications, Flanagan and colleagues (2001) estabilished criteria at both the broad and narrow ability level. If a subtest was classified into a particular broad or narrow ability by 80 percent of its ratings, it was considered to be a measure of that particular ability. The authors made a similar conclusion if a subtest was felt to measure one ability by at least 60 percent of ratings, and no other particular ability received more than 40 percent of the remaining abilities. If a subtest had an agreement of 40 percent or more on two different broad or narrow abilities, it was considered a mixed ability measure. With these evaluation criteria, the authors were able to classify 96 percent of the tests from their pool into broad abilities, and 87 percent of the tests into narrow abilities. If there was no pattern of agreement from the expert ratings, then Flanagan and colleagues classified that subtest themselves.

The Importance of Subtest Task Demands

As Frisby and Parkin (2007) suggest, the CHC classification of subtests appears to be based on both factor analytic and logical analysis data. Tests are classified into broad abilities via factor analysis studies and then further subclassified into narrow abilities through logical analysis which requires "a serious look at the task demands of individual tests" (McGrew, 1997, p. 172). It is possible that such a classification scheme can confound ability and task demand variables and create confusion in subtest classification. Though it is assumed that task demands are related to ability measurement, the extent to which classifications based on ability scores and those based on a logical analysis of task demands are similar is not known. Consider McGrew's (1997) discussion of the Wechsler *Picture Arrangement* test. While expert consensus suggested through logical task analysis that the subtest may measure narrow abilities under *Gf*, joint factor analytic studies demonstrated that the subtest appeared to show small to moderate loadings on *Gv* and *Gc* factors. Similarly, in a study investigating task demand similarities across a wide number of cognitive batteries, Frisby and Parkin (2007) demonstrated how subtests measuring fluid intelligence and visual processing can appear to use similar task demands. In their cluster analysis of 49 different subtests, they demonstrated how tasks such as the WJ-III Cog's *Spatial Relations* or the WISC-*IV's Block Design* test fall in the same cluster as *Matrix Reasoning*. Multidimensional scaling procedures also noted similarities between the WJ-III Cog's *Spatial Relations, Analysis-Synthesis* and *Concept Formation* subtests, which were positioned toward the Performance end of a Verbal/Performance dimension.

Floyd et al's (2005) research on CHC composite score exchangeability may partially reflect this classification problem. In their investigation, they compared *Gf* composite scores between the WJ-III Cog and the DAS, KABC-II and KAIT batteries in samples of children who completed more than one test battery as part of validity studies. The researchers were interested in whether the *Gf* subtests (as well as subtests measuring other broad abilities) from each battery appeared to generate similar scores for the same examinee. After calculating composite scores for each battery, the researchers demonstrated that scores supposedly measuring the same broad ability may be significantly different for approximately 25 percent of their sample. They note that 40 percent of the variance across comparisons of composite scores can be attributed to the combination of random error and systematic error from the interaction between examinees and the test battery. These interactions may include examinee ability level/score characteristics, examinee characteristics/temporal aspects of a test session and examinee characteristics/subtest requirements. Some subtests may not be able to adequately assess the ability of individuals on either extreme of the normal curve due to an inappropriate range of items (i.e. floor or ceiling effects). Temporal aspects of a test session may influence ability measurement due to practice effects, test order, or fatigue. Subtests may also assess extraneous narrow or abilities that are exceptionally well-developed or exceptionally poor within the examinee. Some tasks may also be more appealing to examinees, which may also influence task performance. A direct implication of Floyd et al.'s work, it is possible that a fluid intelligence composite consisting of an inductive reasoning and deductive reasoning measure may produce a very different score in an examinee than a composite that consists of an inductive reasoning and quantitative reasoning subtest. Finally, Floyd et al. stress that subtests' reliability coefficients, their similarity in cultural and linguistic loading, and their use of different types of information processing may all influence exchangeability.

Fluid Intelligence and its Task Demands

Though CHC theory includes a wide variety of broad and narrow abilities within its taxonomy, the present study focuses on fluid intelligence (*Gf*) for a number of reasons, both theoretical and practical. Fluid intelligence represents a key ability in the description and conceptualization of intelligence (Carroll, 1993), and abilities representing fluid intelligence have been included in cognitive batteries since the beginning of the testing movement (Pellegrino & Glaser, 1982). Fluid intelligence can be a useful medium to understand the relationship between subtests task demands and their narrow and broad ability classifications. Fluid intelligence tasks incorporate a wide variety of task demands, yet the narrow abilities classified under *Gf* are not too numerous to investigate on a small scale. Additionally, as Pellegrino and Glaser (1982) note, tasks that describe fluid intelligence (particularly induction) differ in their relationship to each other as a function of both task form and content type.

Through factor analysis, Carroll (1993) isolated a number of Stratum I abilities subsumed under fluid intelligence: induction, general sequential reasoning (deduction), quantitative reasoning, Piagetian reasoning and a processing speed factor. In general, subtests on various intelligence batteries appear to represent the first three abilities, which appear in a broad range of tasks. In fact "the tasks that psychologists have investigated under the heading of reasoning comprise anything but a homogeneous domain, consisting of everything from mental paper folding to cryptanalysis and from rearranging scrambled sentences to planning a nautical course" (Rips, 1984, p. 113).

Induction. As a cognitive ability, induction involves the discovery of an underlying rule, concept or characteristic that pertains to a set of stimuli. Often tasks use analogies, classification, series extrapolation or matrix completions as mechanisms for measurement (Goldman & Pellegrino, 1984). Analogy problems require examinees to complete a statement type, such as cotton:soft :: rock:_____, or 1:3 :: 6:18 :: 2:___. Classification problems may use stimuli such as words, numbers or geometric shapes. Series extrapolation tasks may use letters, numbers or other stimuli. Typically they present the examinee with a progression of stimuli and require the examinee to continue or finish the progression. Matrix tasks consist of puzzles, usually a square with a number of different shapes or stimuli inside, where one piece is missing. Examinees must select an answer to fill in the missing piece from a number of alternative responses. All these tasks require examinees to induce relationships that define associations between stimuli and then to decide a response that fits within those relationships.

Despite differences in task content and presentation, performance on different inductive reasoning tasks appears to be dependent on the same cognitive process (Goldman and Pellegrino, 1984). Studies that have compared multiple ways to measure inductive reasoning have demonstrated striking commonalities in the process involved (Sternberg & Gardner, 1983). However, some researchers note that "correlations for common content tend to be higher than those for common tasks differing in content" (Goldman & Pellegrino, 1984, p. 189). Pellegrino and Glaser (1982) demonstrate that intercorrelations between induction tasks show trends based on both stimuli and task process type through analysis of the *Cognitive Abilities Test* (CAT; Thorndike & Hagen, 1971). The CAT provides reasoning tests that vary in their content (verbal, figural) and task (analogy, classification) dimensions. Correlations between these tasks suggest that a) all induction tasks demonstrate strong relationships with each other, and that b) correlations are stronger for tasks with the same content, rather than for tasks based on the same types of processes. For instance, the *Verbal Analogy* and *Verbal Classification* task correlation is higher (r = .74) than the *Verbal Analogy* and *Figural Analogy* tasks (r = .62). Similarly, *Figural Analogy* and *Figural Classification* correlate higher than *Verbal Classification* and *Figural Classification* (r = .67 vs. r = .57). One reason for these findings may be that fluid intelligence tasks including significant verbal content are likely to tap abilities associated for crystalized intelligence more so than figural content.

In general, success on tasks of induction require the ability to note similarities, dissimilarities and create integrations between test stimuli (Christou & Papageorgiou, 2007). Christou and Papageorgiou suggest that finding similarities may include three types of problems. Class formation problems require examinees to note an attribute that all stimuli share in common. Alternatively, class expansion problems require examinees to decide which attribute one set of stimuli may share with a second stimulus. As a third type, when completing "find the common attribute" problems examinees consider attributes of a number of stimuli in order to ascertain similarities between a specific subset of stimuli. Dissimilarity problems require examinees to induce differences in the attributes of, or the relationships between test stimuli. These tasks may require examinees to decide which particular stimuli does not belong with other stimuli or to note a relationship between stimuli that excludes other stimuli. Integration problems require examinees to consider the attributes of or relationships between a number of stimuli at the same time. For instance, examinees may need to decide if two relationships are equivalent or different, or indicate whether two sets of stimuli consist of similar or different characteristics.

Cross-battery assessment guides (e.g. Flanagan, Ortiz & Alfonso, 2007) note numerous modern cognitive batteries that include tests of induction. The WJ-III Cog's *Concept Formation* task represents an important exemplar, where the examinee is presented with a series of items consisting of circles and squares that vary based on their color, size and number. Any particular item consists of a drawing inside a box and a drawing outside a box. The examinee is required to state a rule that confirms why one drawing is in the box and the other is outside of the box by making comparisons between the drawings. The WISC-IV's *Matrix Reasoning* and *Picture Concepts* also represent tests of induction. For instance, *Picture Concepts* requires examinees to look at a number of rows of pictures, and pick one picture from each row that are similar in some quality.

General sequential reasoning/deduction. General sequential reasoning or deduction involves beginning with known rules and using them to arrive at a solution to a problem. It is a process of generating valid conclusions based on true premises (Johnson-Laird, 1999). Carroll (1993) notes that syllogisms and logical reasoning tasks are representative of this ability. Generally, these tasks present rules and operators that consist of terms such as "some," "all," or "greater than," and "equal to," as well as boolean operators such as "and," "or," and "not." Other tasks may use symbols to describe rules or premises.

Modern cognitive batteries may include tasks of deduction or general sequential reasoning. The WJ-III Cog's *Analysis-Synthesis* test requires examinees to use deduction to solve its items. The examinee is presented with items consisting of a group of two to three colored squares that are associated with a blank square. Each item also includes a key demonstrating how two colored squares are related to another color. The examinee must use the information in the key to decide what color the blank square should be. The *Kaufman Adolescent and Adult Intelligence Test* (KAIT; Kaufman & Kaufman, 1993)'s *Logical Steps* is another illustrative test of general sequential reasoning. In this subtest the examinee uses a set of rules or premises about the relationships between different characters to answer questions about their relative placement in a diagram.

Quantitative Reasoning. Quantitative reasoning reflects the ability to reason with mathematical quantatities. Specifically, the ability involves the application of inductive or deductive reasoning to mathematics including arthimetic, algebra, geometry and calculus (Carroll, 1993). Unlike with inductive or deductive tasks, Carroll does not apply further classification of these tasks. Instead, he suggests that many inductive or deductive reasoning tasks can be made into a quantitative reasoning task, for instance by making the premises in deduction tasks, or the discovery of rules in induction tasks that require the use of mathematical ability. Cognitive batteries such as the WISC-IV, that include subtests like *Arithmetic*, or the WJ-III Cog's *Applied Problems* may tap quantitative reasoning abilities.

Introduction to the Study

The proposed study attempts to determine whether classification of subtests into CHC narrow abilities via their task demands leads to similar conclusions when judging similarity based on ability data (e.g. factor analysis). Results will inform researchers and practioners whether or not task analysis is viable when determining the narrow abilities subtests may measure. It focuses on the classification of fluid intelligence subtests into narrow ability definitions. Narrow abilities represent the third pillar of cross-battery assessment (Flanagan, Ortiz & Alfonso, 1997), and are used to construct and interpret broad ability composites and to tease apart examinees' cognitive strengths and weaknesses. Therefore, an accurate understanding of the narrow abilities measured by subtests is critical to effective assessment. Research identifying subtests' classification into CHC broad and narrow abilities has utilized both factor analytic and expert consensus methodology. Expert consensus studies and task analysis methods can suggest which specific narrow abilities a subtest may measure, and represent a method of content validity (Flanagan, Ortiz & Alfonso, 2007). However, as in the case with the Wechsler *Picture Arrangment* subtest, these methods may confound ability and task demand variables. Expert consensus classification suggested that *Picture Arrangement* may measure fluid intelligence abilities, while factor analysis suggest small to moderate loadings on visual processing and crystalized abilities (McGrew, 1997). Currently, it is unknown exactly how subtest task demands influence ability measurement, but researchers have hypothesized that their information-processing characteristics may play a role in why broad ability composite scores from the same individual may vary across test batteries (Floyd et al., 2005).

To investigate the viability of task-analysis when discerning the narrow ability(ies) a cognitive task may measure, this study compares two different empirical methods of classifying subtests. The first method provides an ability classification of subtests through the factor analysis of subtest scores. The second method classifies subtests via their task demands by asking practitioners to sort subtests into groups based on similarities they perceive in written descriptions of task demands. These sortings are subjected to cluster analysis as a method of classification Using cluster analysis, this method provides an alternative statistical procedure from the agreement-percentages used in previous methods (McGrew, 1997; Flanagan, Ortiz, Alfonso & Mascolo, 2001). It also minimizes the concern over individual research bais evident in the history of task demand classification (Frisby & Parkin, 2007). Lastly, the process used in this method highlights similarities among groups of subtests, because participants compare subtests to each other. Alternatively, previous methods of categorization did not require participants to make comparisons between subtests . In contrast, this study requires participants to make judgments of the relative degree of perceived similarity between subtests.

This investigation uses six subtests hypothesized to measure inductive, deductive and quantitative reasoning abilities as target subtests. These include *Picture Concepts* from the *WISC-IV* and *Concept Formation* from the *WJ-III Cog* as measures of inductive reasoning, *Logical Steps* from the *KAIT* and *Analysis/Synthesis* from the *WJ-III Cog* as measures of deductive reasoning and *Arithmetic* from the *WISC-IV* and *Math Reasoning* from the *WIAT-II* as measures of quantitative reasoning. If these subtests are classified similarly based on task analysis and ability analysis, then it is possible to infer a relationship between the two methods. It would also imply that similarities among subtests in their task demands plays a role in explaining shared variance among subtests observed in factor analytic studies.

Additionally, the proposed study contributes to the literature by providing an ability-focused factor analysis of a CHC broad ability, fluid intelligence. To date, factor analytic research has used a battery-focused methodology by including all subtests of two batteries. As an alternative, the proposed study will include only fluid intelligence subtests from a multiple batteries. By providing a finer classification of fluid intelligence subtests, results from this study should aid practitioners in selecting subtests to ensure adequate construct representation in their batteries and to supplement narrow ability measurement in the face of discrepancies within broad ability composites.

Hypotheses

Hypothesis for the confirmatory factor analysis. The study will use confirmatory factor analysis (CFA) procedures to compare the fit of two models to data provided by six fluid intelligence subtests administered to youth. These models are outlined in Figures 1 and 2. The first model describes a one-factor structure for the six fluid intelligence subtests. It indicates that each subtest will contribute to the measurement of a single latent variable, fluid intelligence. In contrast, the second model suggests that each pair of subtests will measure narrow abilities judge by the literature to be associated with fluid intelligence (e.g. induction, general sequential reasoning/deduction, and quantitative reasoning). In line with CHC theory, it is hypothesized that the three-factor model will best fit the six subtests targeted in this study.

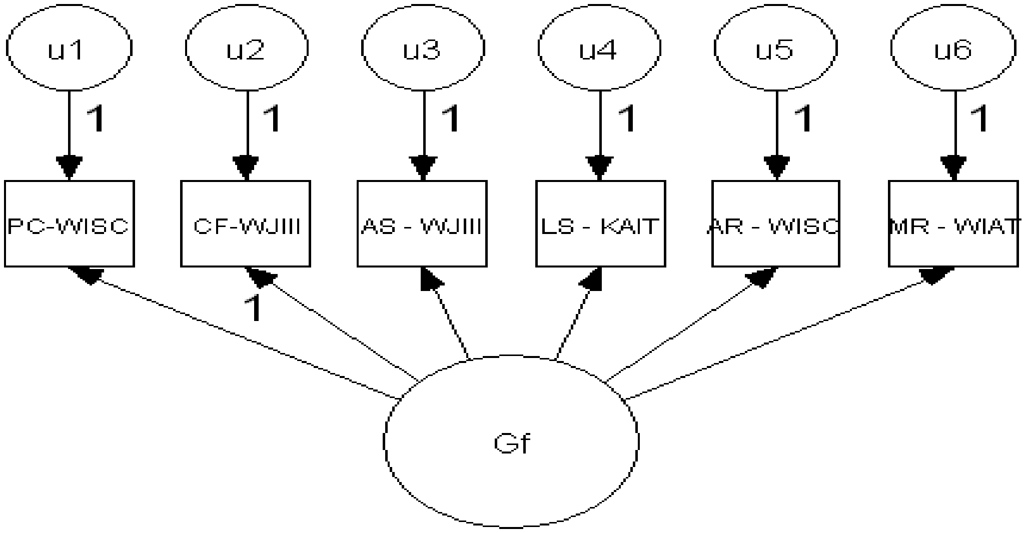
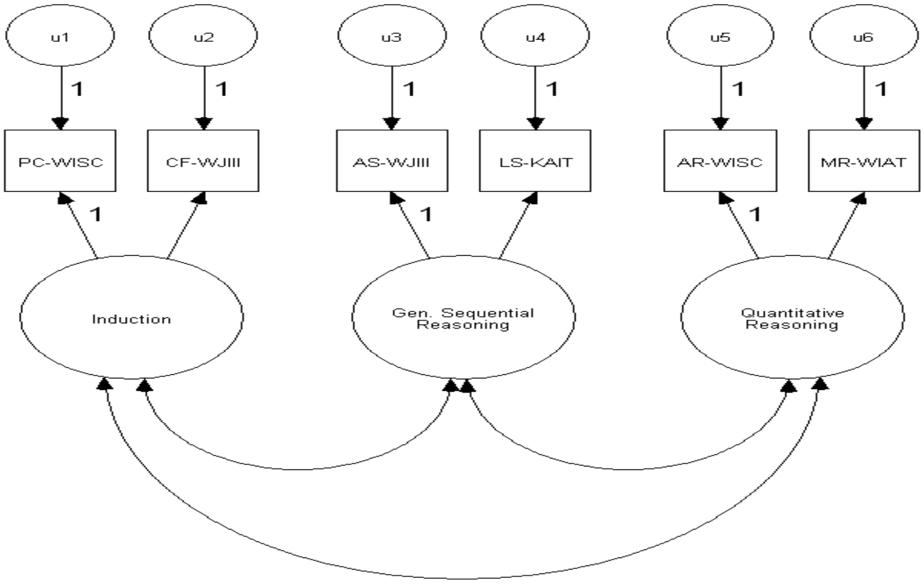


Figure 1. Hypothesized one-factor model for confirmatory factor analysis.



F*igure 2.* Hypothesized three-factor model for confirmatory factor analysis.

Hypotheses for the task analysis. In this part of the study, participants will sort brief subtest descriptions into similar groups based on their functional task characteristics. Subtests include tasks that involve reasoning skills (e.g. *Gf*), but also math skills, processing speed and visual-spatial abilities, as “distractors.” Distractors are included to ensure that participants are attending to cognitive processes implied in the cognitive subtests, and not the superficial characteristics of the task. It is hypothesized that induction tests will be sorted into the same group, as will deduction and quantiative reasoning tests. However, likely the mathematics knowledge tests will cluster with the quantitative reasoning tests, despite their different functional requirements.

Hypotheses integrating both methods. If these methods provide similar results, than the factor analysis will indicate that a three-factor model bests fits the ability data. Similarly, the cluster analysis will sort the six target subtests into separate clusters.

CHAPTER THREE: METHODS

Approval by Human Subjects Committee

The investigator of this study completed ethics training required by the Institutional Review Board (IRB) at the University of Missouri (MU). The project, and its associated materials, was submitted to, and approved by the MU IRB.

Participants

Phase 1 Ability Analysis. The participants in this investigation were 63 respondents between the ages of 11 and 16 (mean age = 12 years 9 months). The age range is reflective of the minimum and maximum ages included in the norming samples of some of the cognitive instruments used in this study. Approximately 68.3 percent of the respondents in this sample were female. Ethnicity was judged informally by the investigator; 84.1 percent, were white, 12.7 percent were African American, and 3.2 percent were Latina/o.

Phase 2 Task Analysis. Respondents consisted of a convenience sample of 43 individuals, employed in Missouri school districts, who have extensive experience with psychoeducational assessment. Out of the 43 individuals who participated, 11 individuals provided incomplete data, and as a result, their responses could not be included as part of the cluster analyses used to analyze data. Thus, the final sample included 32 professionals in school psychology.

As part of their participation, respondents were asked to provide basic demographic information, and also data about their professional experiences.

Specifically, respondents provded their age, race and gender. They also indicated their years of professional experience, their experience with specific psychometric tests, and rated their familiarity with test intepretation through CHC theory on a 1-5 point likert scale. This information is summarized in Table 1, and a copy of the questionnaire is included in Appendix B. To ensure that there were no differences between individuals who completed and did not complete the sorting task, a chi-square (χ^2)test of independence was used to report whether there was a difference in the frequency that each group endorsed experience with each psychometric test. The results, listed in Table 1, were interpreted using a conservative ($\alpha = .01$) alpha level to correct for type I errors due to the numerous comparisons. Results indicated that there were no differences between the sample of individuals that provided complete data to the task analysis and the sample of individuals who provided incomplete data in terms of their use of specific cognitive tests. Also listed in Table 1, independent samples *t*-tests indicate that there is no difference in age, $t(39) = .201$, $p = .842$, or years of experience, $t(27) = -.492$, $p = .627$ between the two groups. The groups also endorsed a similar degree of CHC test interpretation in their work, $t(40) = -.665$, $p = .51$. However, the individuals who did not provide complete sorting data collectively endorsed a higher knowledge of CHC theory compared to the group that provided complete sorting data, $t(33.52) = -2.41$, $p = .02$. Due to signficant differences in variance between groups, as assessed through Levene's Test of Equality of Variance, a corrected t-test was interpreted.

Table 1.

Card Sorting Sample Demographic Information

	Complete Sort	Incomplete Sort
	(N= 32)	(N = 11)
	Mean (SD)	Mean (SD)
Age (Years)	42.1 (12.06)	41 (12.97)
Years Paid	11.67 (9.96)	14 (7.58)
	N(%)	
Position		
Intern	3 (9.4)	0 (0)
Psychology Examiner	6 (18.8)	3 (27.3)
School Psychologist	18 (56.3)	8 (72.7)
Licensed Psychologist	4 (12.5)	0 (0)
Diagnostician	1 (3.1)	0 (0)
Test Experience*		
Bayley I	11 (34)	4 (36.4)
Bayley II	4 (12.5)	2 (18.2)
CAS	3 (9.4)	0 (0)
CTONI	20 (62.5)	7 (63.3)
DAS	8 (25.0)	4 (36.4)
DAS-II	13 (40.6)	2 (18.2)
DTLA-3	3 (9.4)	0 (0)
DTLA-4	2 (6.3)	0 (0)
KABC	16 (50.0)	6 (54.5)
KABC-II	12 (37.5)	6 (54.5)
KAIT	2 (6.3)	0 (0)
K-BIT	9 (28.1)	3 (27.3)
Leiter	23 (71.9)	8 (72.7)
RIAS	9 (28.1)	0 (0)
SIT-R	5 (15.6)	3 (27.3)
SB-IV	19 (59.4)	9 (81.8)
SB-V	23 (71.9)	9 (81.8)
UNIT	23 (71.9)	7 (63.6)
WAIS-III	22 (68.8)	8 (72.7)
WISC-III	27 (84.4)	9 (81.8)
WISC-IV	31 (96.9)	11 (100)
WJ-R	13 (40.6)	6 (54.5)
WJ-III Cog	22 (68.8)	8 (72.7)
WPPSI-R	18 (56.3)	3 (27.3)
WPPSI-III	21 (65.6)	6 (54.5)

	Mean (SD)	Mean (SD)
CHC (Likert Scale)		
CHC Knowledge	3.16 (1.04)	3.7 (0.48)*
Use CHC in Practice	2.60 (0.87)	2.8 (0.79)

**signficant, p* < .05 - percentages reflect the number of participants that endorsed having experience administering a specific test - CHC knowledge was assessed on likert scales with a range of 1-5.

Instruments (Materials)

Phase 1 Ability Analysis. Subtests were selected as measures of *Gf* subtests based on classifications identified in previous literature (Buckhalt, McGhee, & Ehrler, 2001; Flanagan & McGrew, 1998; Flanagan, Ortiz & Alfonso, 2007; Flanagan, Ortiz, Alfonso & Mascolo, 2001; Keith, Kranzler & Flanagan, 2001; Phelps et.al., 2005; Sanders et. al., 2007; Woodcock, 1990) and represent inductive, deductive and quantitative reasoning abilities. These abilities were measured by subtests from the *Wechsler Intelligence Scale for Children, Fourth Edition,* the *Kaufman Adolescent and Adult Intelligence Test*, the *Woodcock Johnson Test of Cognitive Abilities, Third Edition,* and the *Wechsler Individual Achievement Test, Second Edition.* Inductive tests include the WISC-IV *Picture Concepts* and the WJ-III Cog *Concept Formation*. *Analysis/Synthesis*, also from the WJ-III Cog and *Logical Steps* from the KAIT represent deductive subtests. *Arithmetic* (WISC-IV) and the WIAT-II's *Mathematical Reasoning* subtest reflect quantitative reasoning measures. These tests were selected specifically for their varied task demands in comparison to one another. For instance, the KAIT, though an older measure (normed in 1993), provides an important measure of deductive reasoning (Flanagan & McGrew, 1998) that appears more unique in terms of task demands placed on the examinee than the demands from more current cognitive batteries. In comparison, *Visual Coding*, is a subtest from the Leiter-R*,* whose task demands appear too similar to the *Analysis-Synthesis* subtest from the WJ-III Cog. Thus, not only has the KAIT *Logical Steps* subtest been empirically classified as measuring general sequential reasoning, it would be a better test to use when comparing these results to the phase 2 (task demands classification study) because it provides more task-diversity to the six subtests used in this phase of the study.

Phase 2 Task Analysis. Rationale for Subtest Selection. Phase II of the study classifies subtests based on their task demands. Authors of cross-battery assessment manuals (Flanagan, McGrew & Ortiz, 2000; Flanagan, Ortiz & Alfonso, 2007; McGrew & Flanagan, 1998) and CHC theory researchers (Flanagan & McGrew, 1998; Keith, Fine, Taub, Reynolds & Kranzler, 2006; McGrew, 1997; Phelps, McGrew, Knopik & Ford, 2005; Reynolds et. al., 2007; Sanders, McIntosh, Dunham, Rothlisberg & Finch, 2007; Taub & McGrew, 2004) have created categorizations of subtests at the broad and narrow ability level. These materials were used to create a list of subtests that are hypothesized to be primary measures of fluid intelligence to be included in this part of the study. ”Distractor” subtests (subtest descriptions other than the “target” subtests) were included to a) provide enough stimuli to create meaningful clusters, and b) investigate whether participants were making similarity judgments based on the surface level characteristics of tasks, or the underlying processes they require for completion. In short, this enables the researcher to determine how participants classify the target subtests in the context of other subtests. Generally, subtests that appear to be partial measures of fluid intelligence abilities, but also measure other cognitive abilities (e.g. *Rover* from the KABC-II, *Planning* from theWJ-III Cog) were not included in the classification task. This requirement is essential to make sure the study corresponds to important CHC principles, specifically, avoiding “mixed” tests to ease interpretation issues (Flanagan, Ortiz & Alfonso, 2007).

However, there will be two exceptions to this rule. First, the study used the WISC-IV *Arithmetic* subtest even though researchers have argued about which ability(ies) this test may measure (Keith et. al., 2006). However, recent confirmatory factor analyses of the WISC-IV suggest that Arithmetic may be best interpreted as a primary measure of fluid intelligence (Keith et. al.) and therefore should be included in this study. Second the study included a number of “distractor” subtest descriptions. These subtests appear to have surface level features that make them seem similar to subtests that primarily measure fluid intelligence, but they nevertheless are classified as measures of other types of Stratum II abilities.

Subtests from the test batteries analyzed by the literature were further paired down by removing subtests that appeared to have duplicate task demands with other subtests. For instance, the DAS-II (*Picture Similarities*) and the Leiter-R (*Classification*) both include subtests that require participants to match a target picture to another picture within a group of target stimuli. Thus, only one of these subtests was be included as part of this study.

There was one exception to this rule. *Applied Problems*, a mathematics reasoning test from the WJ-III Cog, and *Math Reasoning*, a subtest from the WIAT-II, though similar in their task demands, will both be included as a part of this study as a validity check. The purpose of this procedure is to provide an indirect assessment of participants’ vigilance through out the task. Due to the similarities of these two tests, if participants were vigilant in completing their sorting, these two tests should be positioned closely to each other.

Creation of Subtest Descriptions. Subtest descriptions were written to highlight their task demands, but did not explicitly mention specific cognitive abilities that may tapped as they are completed. All subtest descriptions were created with the same language, and used the same structure to minimize any influence a writing style may have on how participants may interpret the description.

Subtests Hypothesized to Be Measures of Induction. Selected subtests that measure inductive reasoning processes were selected from the *Woodcock-Johnson Test of Cognitive Abilities, Third Edition* (e.g. *Concept Formation*), the *Weschler Intelligence Scale for Children, Fourth Edition* (e.g. *Picture Concepts*), the *Kaufman Assessment Battery for Children, Second Edition* (e.g. *Pattern Reasoning*), the *Kaufman Adolescent and Adult Intelligence Test* (e.g. *Mystery Codes*), and the *Leiter International Performance Scale – Revised* (e.g. *Classification, Repeat Patterns*). Descriptions of these subtests' task demands are provided in Appendix A.

Subtests Hypothesized to Measures of General Sequential Reasoning. Selected subtests that appear to measure general sequential or deductive reasoning were included from the *Woodcock-Johnson Test of Cognitive Abilities, Third Edition* (e.g. *Analysis/Synthesis*), the *Kaufman Adolescent and Adult Intelligence Test* (e.g. *Logical Steps*), and the *Leiter International Performance Scale – Revised* (e.g. *Picture Context, Visual Coding*). Descriptions of these tasks are located in Appendix A.

Subtests Hypothesized to Measure Quantitative Reasoning. Selected subtests that appear to measure quantiative reasoning were included from the *Woodcock Johnson Test* *of Cognitive Abilities, Third Edition Diagnostic Supplement* (e.g. *Number Matrices, Number Series*), the *Wechsler Intelligence Scale for Children, Fourth Edition* (e.g. *Arithmetic*), the *Differential Abilities Scale, Second Edition* (e.g. *Sequential & Quantitative Reasoning*), and the *Wechsler Individual Achievement Test, Second Edition* (Wechsler, 2001) (e.g. *Math Reasoning*) and the *Woodcock Johnson Test of Achievement, Third Edition* (e.g. *Applied Problems*). Appendix B includes descriptions of these subtests.

Subtests Hypothesized to Measure More than One Gf Narrow Ability. Subtests from the *Kaufman Assessment Battery for Children, Second Edition* (e.g. *Story Completion*), the *Wechsler Intelligence Scale for Children, Fourth Edition* (e.g. *Matrix Reasoning*) may measure more than one narrow, fluid intelligence ability were also included in this study. These subtests were included because many of their characteristics were similar to some of the six target subtests, and also because they may measure both inductive and deductive reasoning – their descriptions are located in Appendix A.

Subtests Included as Distractors. A number of additional subtests that have been classified by cross-battery advocates as measures of other CHC abilities were also included in the card sorting task. These tests were included because they share surface features with many of the fluid intelligence subtests discussed earlier. Their inclusion will test whether participants sort subtests by the cognitive processes they appear to measure, as opposed to sorting by functional characteristics and surface features.

Many of these distractor subtests are thought to measure visual-spatial abilities (e.g. Flangan, Alfonso & Ortiz, 2007). Many visual-spatial tests appear to share similarities in common with fluid intelligence subtests. For instance, Frisby and Parkin (2007) showed that both Gf subtests and Gv subtest appear to be characterized as “performance” subtests. Other subtests were chosen for the different ways they use numbers as stimuli to assess different cognitive abilities. *Digit-Span*, from the WISC-IV, uses numbers to assess memory span abilities. Alternatively, *Calculation* (WJ-III Achievement) and *Numerical Operations* (WIAT-II) use numbers as math problems to assess knowledge of mathematical principles. Some subtests were included because they have a key or a code as part of their task demands, a hallmark of deductive reasoning subtests. *Digit-Symbol Coding* (WISC-IV), and *Planned Codes* (CAS) require examinees to use a code to complete their tasks, and measure aspects of processing-speed. *Visual-Auditory Learning* (WJ-III Cog) teaches examinees a code to assess how they are able to associate two types of information together. If participants sort subtests merely based on surface feature similarities (and not in terms of cognitive processing requirements) these distractor subtests may (incorrectly) appear in clusters with the fluid intelligence subtests. However, if participants sort the subtests more by their processing requirements, these subtests may appear in alternative clusters. Descriptions of these subtests are located in Appendix A.

Materials Given to Participants. To complete this part of the study, participants were given a set of 32 index cards with subtest descriptions printed on them (see Appendix A), and an answer form to record their responses (see Appendix B).

Instructions. Participants received oral and written instructions to read all cards and then sort them into piles of similarities based on the abilities described through the written task descriptions. A copy of the instructions is given in Appendix B. They were told to make at least 3 piles of cards and to place at least 2 cards in each pile. These requirements were necessary for analysis purposes. Since the six target subtests were hypothesized to reflect three different narrow abilities, requiring participants to sort test descriptions into a minimum of three piles was necessary to determine if participants readily distinguish between inductive, deductive and quantitative reasoning. Requiring at least two cards to be included in one pile or category was necessary, because similarity judgments between cards were the unit of analysis. If only one card was included in a pile, no similarity data would be generated in that category. After sorting the cards, participants recorded their results on a response sheet (see Appendix B).

Procedure

Phase 1 Ability Analysis. Participants were recruited through a summer school program at a middle school in rural, central Missouri. Recruitment flyers and consent forms were sent home to parents which explained participation requirements and the pros and cons of participation (see appendix C). Any student who returned their consent form was eligible to participate. All participants were administered the six subtests, in random order, by the investigator at their school program. Each participant received a $5 gift certificate to McDonald's as a token of appreciation for their participation.

Phase 2 Task Analysis. Participants were recruited from school districts in the state of Missouri. School districts' directors of psychological services were contacted by the investigator to inquire about opportunities for data collection. With the director's permission, the investigator collected data in a group format, during a department meeting at three different sites. Participants were given a consent form along with a response sheet and a stack of index cards that contained printed subtest descriptions. If an individual wished to participate, they completed the sorting task, which required approximately half an hour. At the end of the task, participants were entered into a drawing for a gift certificate, as a token of appreciation.

Data Analysis

Phase 1 Ability Analysis. The six subtests used in this part of the investigation are interpreted through different measurement scales in clinical practice. Scores from *Arithmetic, Picture Concepts* and *Logical Steps* are interpreted by using a mean of 10 and a standard deviation of 3. *Math Reasoning, Concept Formation* and *Analysis-Synthesis* use a mean of 100 and a standard deviation of 15. Thus, in order to ease interpretation, before statistical analysis, the *Arithmetic, Picture Concepts, and Logical Steps* subtest scores were converted to a mean of 100 and a standard deviation of 15 with the following formula: ((X – 10)/3) * 15 + 100. This formula converts their scores to Z scores, and then converts the Z scores to a scaled score having a mean of 100 and a standard deviation of 15.

As described in the review of relevant literature, past research has outlined the hypothesized factor structure for the six subtests in this phase of the study in accordance with CHC theory. Confirmatory factor analysis (CFA) procedures were used for model testing. Often consdered more theory driven that exploratory factor analysis (EFA; Keith, 2005), CFA allows for the creation and testing of hypotheses by the comparison of competing models. Two models were tested: 1) a one-factor model that specified all subtests loading onto the same factor (posited to be a *Gf*, fluid intelligence factor), and 2) a three factor model that specified each pair of subtests loading together, as would be predicted by CHC theory. A χ^2 test was used as an initial test of each model's fit. The χ^2

test evaluates whether the covariance matrix in the data set is significantly different from the covariance matrix implied by the model (Keith). Thus, a significant χ^2 test indicates that the modeled relationships reflect a poor fit to the data. It is important to note that the χ^2 test is sensitive to sample size (Brown, 2006; Keith, 2005; Klem, 2000). When used with a large sample size, a χ^2 test may suggest that a model's covariance matrix is significantly different from the data's covariance matrix, even when it is not. Likewise, with a small sample size, the χ^2 test may indicate an excellent fit even if one does not exist.

Due to the limitations associated with the χ^2 test, a number of goodness of fit statistics were calculated to assist in determining the degree to which these models fit the results of the subtest measurement (Brown, 2006; Hu & Bentler, 1999; Keith, 2005). These fit statistics were chosen because they reflect absolute, comparitive, and parsimony correction fit statistics. Further, they reflect some of the more popular in the research base, and perform well in Monte Carlo research (Brown, 2006). The Tucker-Lewis index (TLI) and the comparitive fit index (CFI) were calculated to compare each model to a "null" model where all variables are assumed to be unrelated. TLI and CFI values over .95 suggest excellent model fits, and values greater than .90 suggest adequate fit. The root mean square error of approximation (RMSEA) was assessed to measure the model's approximate fit to the data (in comparison to χ^2 evaluation of an *exact* fit). Alternatively, the SRMR represents the average or standardized discrepancy between the correlations presented in the correlation matrix, and the correlations predicted by the hypothesized models (Brown, 2006; Keith, 2005). Low values for the SRMR and RMSEA (below .06) suggest that there is little descrepency between the two matrices and indicate better fits.

Akaikes Information Criteria (AIC) and Bayesian Information Criterion (BIC) were also used to inspect the relative fit of one model to the other. Values from these fit statistics are interpreted relative to their results from different models. When comparing models, the model with the lower AIC and BIC value represents the model of better fit.

Phase 2 Task Analysis. Each participant's responses were converted into a 32 x 32 similarity matrix. Each value in a cell represented the total number of times a subtest pair were sorted into the same pile. For example, if a participant sorted subtest 5 and subtest 28 into the same category, a 1 would be entered into the (5, 28) matrix cell. After a matrix had been made for each participant in the study, all values in each cell were summed to create a *grand* 32 x 32 matrix. In this final matrix, if the cell (10, 19) contained a score of 0, it would mean that subtest 10 and subtest 19 were not sorted into the same category by *any* participant in the study. Alternatively, higher values within a cell means that more participants sorted the subtests together, and hence they would have a higher pairwise similarity rating.

These values provide similarity distance scores for use in cluster analyses specifying a range of solutions containing three to seven clusters. Cluster analysis is an appropriate technique for this investigation because it classifies data into groups based on investigator specifications and will create a specified number of homogenous groups of subtests. Thus, the abilities measured by subtests in the same cluster, as assessed by subtest task demands, are considered to be similar to each other.

SPSS' QUICK SORT, a non-hierarchical analysis was used, because it is not agglomerative in generating solutions. This means that a six cluster solution, is not merely the result of combining two clusters from a seven cluster solution, but rather a solution that creates the best six cluster solution, cluster membership will be free to vary based on the specified target solution. (Hair & Black, 2000). SPSS' QUICK SORT algorithym is considered a sequential threshold non-hierarchical procedure (Hair & Black). It creates initial seed points based a random observation in the data set, and then includes all observations within a certain different with it as a cluster. Next, it selects a second random seed point and continues until all specified clusters are created.

Data was subjected to analyses specifying three through seven cluster solutions. Three was used as the initial solution because it reflects the minimum amount of clusters that participants were asked to create. Seven was used as the highest number of clusters, as it reflects the largest hypothesized amount of cluster that participants may create while completing the sorting task.

CHAPTER FOUR: RESULTS

Phase I: Classification Based on Factor Analysis

Data Screening. Data screening was conducted in SPSS, version 11.5 (2002) to check whether data met assumptions for multivariate analysis, including normality, linearity and homoscedacticity. The assumption of normality means that the distribution of scores within a variable approximately reflect a normal curve. Univariate normality was assessed for each subtest's distribution by comparing its kurtosis and skewness to a null hypothesis of zero with a conventional, albeit conservative alpha level of .001 (e.g. Z = 3.29; Tabachnick & Fidell, 2007). As indicated in Table 2, not only were skewness and kurtosis levels for all variables non-significant with the conventional alpha level, all were within one standard deviation of 0 (e.g. 1.96), suggesting that all distributions were close to appropriate levels of normality. The assumption of homoscedacticity stresses that the variability in scores from one variable is approximately the same as variability in scores from another variable. Linearity assumes a stright-line relationship between two variables. Both homoscedacticity and linearity were assessed by inspecting bivariate scatterplots between all combiniation of variables. There were no variable pairs that demonstrated significant violations of heteroscedasticity and all demonstrated linear relationships.

Analyses were run using Mplus (Muthen & Muthen, 2007) with maximum liklihood modeling (ML), using the raw scores for input. An important assumption of ML modeling is that variables conform to a multivariate normal distribution (Kline, 2004). Multivariate kurtosis was assessed with Mardia's Coefficient, using the SPSS macro provided by DeCarlo (1997). Results indicated that multivariate normality was within acceptable limits. Multivariate outliers were assessed through the calculation of Mahalanobis distance. No outliers were detected using a conservative alpha level (.001), as recommended by Tabachnick and Fidell (2007).

Table 2.

Normality Testing

	Skewness	Kurtosis
Picture Concept	1.7	.30
Concept Formation	.35	.77
Analysis/Synthesis	1.68	1.74
Logical Steps	.08	.77
Math Reasoning	1.83	.34
Arithmetic	1.30	1.35

Values higher than 1.96 represent p < .05; Values higher than 3.29 represent p < .001

The data set was also assessed for bivariate and multivariate multicollinearity through an inspection of bivariate correlations, and through collinearity diagnostics available through SPSS. Multicollinearity can occur in a correlation matrix when variables are too highly correlated with each other and can result in the specification of improper solutions during CFA analyses (Brown, 2006). Table 3 indicates that there are a number of correlations approaching or above .70, though these values do not suggest the existance of bivariate multicollinearity in the sample data. Similarly, collinearity diagnostics do not indicate multicollinearity in the data set through multivariate correlations. Condition index values, measures of the dependency of one variable on others in the data set, are associated with instability in the standard error of parameter estimates for a given variable (Tabachnick & Fidell, 2007). When condition index values are above 30 and the variance proportion of two or more variables load highly on that index (e.g. .50 or higher), it indicates a high level of multicollinearity (Tabachnick & Fidell, 2007). The condition indexes for the data set are listed in Table 4. Both dimension six and seven demonstrate condition indexes over 30. In this analysis, dimensions reflect eigenvalues, measures of variance within a given matrix (Tabachnick & Fidell, 2007). The *Analysis/Synthesis* and *Math Reasoning* subtests both load very highly on dimension seven. However, it is noteworthy that this particular dimension has a value of 0, which means it is not associated with variance in the data matrix. Thus, even though dimension seven has a high condition index and is highly associated with two different variables, results of the diagnostics do not indicate significant multicollinearity.

Table 3.

Descriptive Statistics and Correlations for Gf Subtests for n = 63 participants.

	Mean (SD)	PC	CF	AS	LS	MR	AR
Pic. Con (PC)	98.17 (17.35)						
Con.Form (CF)	100.41 (15.75)	.49					
An/Syn (AS)	97.33 (11.74)	.35	.60				
Log.Steps (LS)	104.44 (15.48)	.43	.58	.40			
MathRea (MR)	96.75 (16.73)	.44	.72	.66	.67		
Arith (AR)	97.33 (11.74)	.32	.57	.50	.57	.70	

All correlations significant at .01 level

Table 4.

Multicollinearity diagnositics with six Gf subtests

			Variance Proportions					
Dimension	Eigen Value	Cond. Index	PC	CF	AS	LS	MR	AR
1	6.93	1.000	.00	.00	.00	.00	.00	.00
2	.02	17.89	.64	.00	.00	.00	.04	.12
3	.02	20.93	.20	.02	.07	.01	.05	.01
4	.01	24.98	.02	.29	.11	.17	.03	.25
5	.01	26.50	.11	.01	.02	.51	.01	.51
6	.01	31.42	.03	.68	.18	.00	.29	.08

7	.00	41.84	.00	.00	.63	.30	.57	.03

Confirmatory Factor Analysis of the Three-Factor Model. The three-factor model specifying an induction (*Picture Concepts* and *Concept Formation*), deduction (*Analysis/Synthesis* and *Logical Steps*), and quantitative reasoning (*Math Reasoning* and *Arithmetic*) factor (see Figure 2) was analyzed in Mplus (Muthen & Muthen, 2007) with maximum liklihood modeling (ML), using the standard scores for input. Though the model converged normally, the latent variable covariance matrix was not positive definite. Specifically, the relationship between the deduction factor with both the induction and quantitative reasoning factor was estimated to be greater than 1. Thus, results can not be interpreted appropriately. This improper solution is most likely a result of the low sample size used in the analysis. A number of attempts were made to respecify the model. These included contraining the factor loadings of each pair of indicators on each latent variable to be equal, and constraining the error variance associated with the deducation latent factor to 0. These attempts were unsuccessful in fixing the model.

However, it was still possible to compare a two factor model, consisting of induction and quantitative reasoning factors to a one factor model (e.g. Gf). The rest of these analyses, compared a two factor model to the one factor model using the *Picture Concept, Concept Formation, Arithmetic* and *Math Reasoning* subtests. *Logical Steps* and *Analysis/Synthesis* were deleted from the analysis, because their latent factor was associated with the CFA's improper solution.

Confirmatory Factor Analysis of the One Factor and Two Factor Models. A reassessment of multicollinearity in a data set only including *Math Reasoning, Arithmetic, Picture Concepts* and *Arithmetic* subtests is listed in Table 6. Though both *Matrix Reasoning* and *Concept Formation* load heavily on the fifth dimension, since the dimension's Condition Index is not greater than 30, it can be reasoned that there is no significant multicollinearity in the new variable set.

Table 5.

Multicollinearity diagnostics with four subtests

			Variance Proportions			
Dimension	Eig. Value	Cond. Idx	PC	CF	MR	AR
1	4.95	1.00	.00	.00	.00	.00
2	.021	15.21	.61	.01	.06	.15
3	.014	18.49	.20	.03	.08	.00
4	.010	22.15	.19	.46	.02	.58
5	.01	27.79	.00	.51	.83	.26

Two new models were specified based on subtests measuring inductive and math reasoning. Figures 3 and 4 illustrate the model and provide the factor loadings for each subtest. Model fit statistics are presented in Table 6. The Gf, one-factor model's χ^2 value is non-significant, suggesting that it's estimated covariance model and the covariance model provided by the actual data are not statistically different from each other. Other fit statistics (CFI, TFI, SMRS and RMSEA) also indicate an excellent fit to the data. CFI and TFI values are greater than .95, the RMSEA is closed to .06, while the SMRS is significantly lower than .06.

CFA results indicate that a two-factor model provides an excellent fit to the sample data as well (see Table 6). The model's χ^2 value CFI, TLI, SMR and RMSEA values are all in the range of excellent fit (Brown, 2006; Hu & Bentler, 1999; Keith, 2005). Moreover, they indicate that the two-factor model is a better fit than the one-factor, Gf model. The CFI value of 1.00 represents the maximum value possible for the statistic. The TLI value, 1.05, is interpreted similar to the CFI. However, it is possible for its values to exceed 1.0 (Brown, 2006). Additional statistics included in the table, the Akaike information criterion (AIC) and the Bayes information critierion (BIC) are measures of relative fit (Keith, 2005). These statistics are interpreted relatively to other models, where smaller values indicate better fits. These statistics provide confliciting results; while the AIC suggests the two factor model is a better fit, the BIC favors the one-factor model. Though both models represent excellent data fits, the two-factor model appears to be the better fitting model. The CFI, TFI, RMSEA and SMRS fits favor it, and in general, it's factor loadings are larger, while its error variances are smaller, compared to the one-factor model.

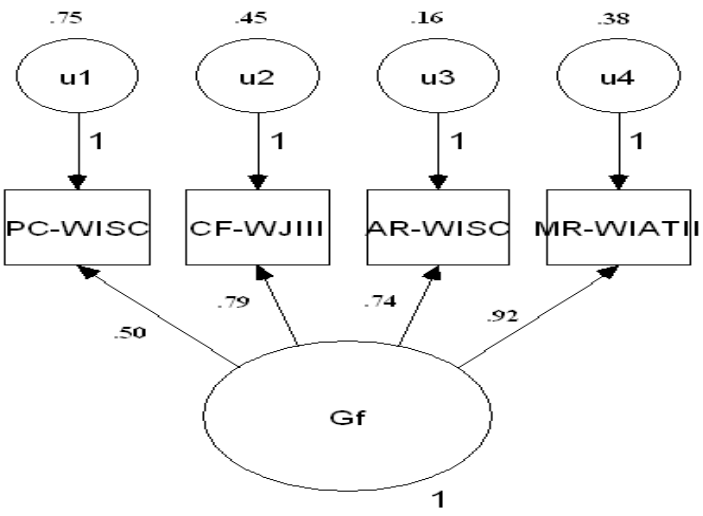


Figure 3. One-factor model with four subtests, including factor loading.

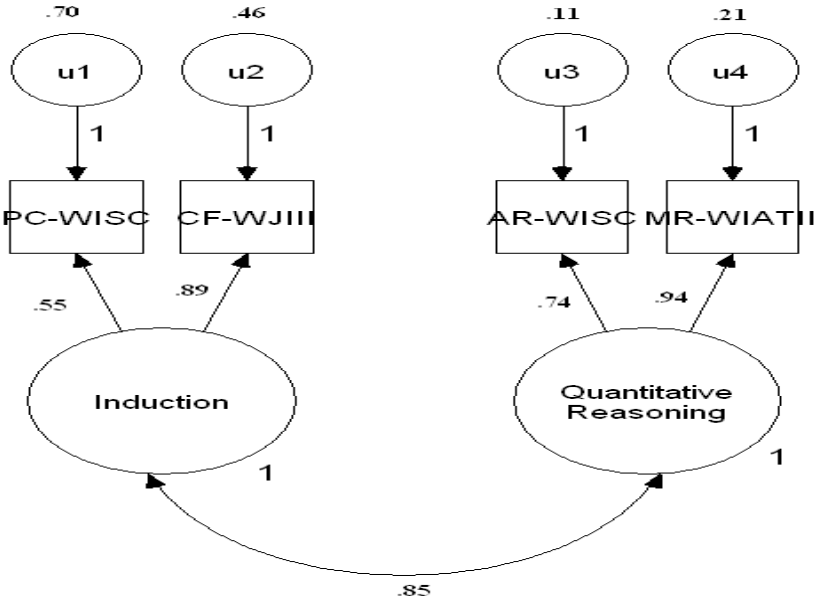


Figure 4. Two-factor model with four subtests, including factor loadings.

Table 6.

Fit statistics comparing one-factor and two-factor models

Model	χ^2	*Df*	*P*	CFI	TFI	RMSEA	SRMR	AIC	BIC
Gf	2.69	2	.26	.993	.980	.07	.03	2042.84	2068.56
I + QR	.10	1	.76	1.00	1.05	.00	.01	2042.25	2070.11

CFI: Comparative Fit Index; TLI: Tucker Lewis Index; RMSEA: Root Mean Square Error of Approximation; SRMR: Standard Root Mean Square Residual; AIC: Akaike's Information Criterion; BIC: Bayesian Information Criterion.

Phase II: Classification Based on Functional Task Analysis

The grand similarity matrix, created by summing the cells of each participant's responses, was subjected to 3 through 7 means nonhierarchical cluster analysis through SPSS' QUICK SORT algorithm. The results of these sortings are provided in Table 7, where each subtest's cluster membership can be compared across 3 through 7 cluster solutions. The six target subtests and validity check subtest are highlighted in bold. The hypothesized deductive tasks, *Analysis/Synthesis* and *Logical Steps* clustered together when 5 clusters were specified. They continued to cluster together in the 6 and 7 cluster solution. The inductive tasks, *Picture Concepts* and *Concept Formation* clustered together in the initial 3 cluster solution and continued to cluster together in the other solutions. *Math Reasoning* and *Arthimetic*, the quantitative reasoning subtests, clustered together in all solutions except the 7 cluster solution. Results indicated that five clusters was the minimum number of clusters required to determine when these three sets of subtests will cluster together.

Table 7

Subtest Cluster Membership By 3,4,5,6, and 7 Cluster Solution

Subtest Name	Cluster Solution				
	3	4	5	6	7
Analysis/Synthesis (Deduction)	**2**	**4**	**1**	**1**	**1**
Logical Steps (Deduction)	**1**	**2**	**1**	**1**	**1**
Picture Concepts (Induction	**2**	**4**	**4**	**2**	**7**
Concept Formation (Induction)	**2**	**4**	**4**	**2**	**7**
Arithmetic (Quant. Reasoning)	**3**	**3**	**3**	**3**	**3**
Math Reasoning (Quant. Reas)	**3**	**3**	**3**	**3**	**2**
Applied Problems (Validity)	**3**	**1**	**2**	**4**	**4**
Picture Arrangement	1	2	1	1	1
Digit-Sym Coding	1	2	5	5	5
Numerical Operations	3	3	3	3	3
Word Reasoning	1	2	5	6	6
Mystery Codes	2	4	1	1	1
Vis.Aud. Learning	1	2	5	5	5
Classification	2	4	4	2	7
Planned Codes	1	2	1	5	5
Digit Span	1	2	5	5	5
Odd Item Out	2	4	4	2	7
Matrix Reasoning	2	4	4	2	7
Similarities	1	2	1	1	1
Story Completion	2	4	1	1	1
Spatial Relations	2	4	4	2	7
Repeated Patterns	2	4	4	2	7
Seq. Quant. Reas.	2	4	4	2	7
Symbolic Memory	1	2	5	5	5
Pattern Reasoning	2	4	4	2	7
Nonverbal Memory	1	2	5	5	5
Calculation	3	3	3	3	3
Visual Coding	2	4	4	2	7
Verbal Spatial Rel	1	2	1	6	6
Number Matrices	2	4	4	2	7
Picture Completion	2	4	1	1	7
Number Series	2	4	1	1	1

The contents of each cluster within each solution are described in Tables 8 through 12. The three cluster solution outlined one cluster mixed with fluid reasoning and distractor subtests, and second cluster containing mostly (though not exclusively) memory-oriented subtests, and a third cluster containing only math related tests. However, it is interesting to note that the third cluster did not contain all possible math subtests. *Number Series* and *Number Matrices* were both placed in the second cluster. When a four cluster solution was specified, the addition of a fourth cluster only moved the *Applied Problems* subtest to its own cluster, all cluster membership of all other subtest descriptions remained the same as the previous solution. The fivecluster solution moved the memory-oriented subtests into their own cluster, along with *Coding* and *Word Reasoning*. This analysis also reorganized the non-math subtests from the four cluster solution. The first cluster contained most of the deductive reasoning tasks and tasks that required the reorganization of pictures, while the other reflected many inductive reasoning tests and subtests with abstract stimuli. These are merely generalizations, as each cluster contains some tests that do not fit neatly into these descriptions. The 6 cluster solution created an additional verbal-related cluster, though it did not contain all of the heavily verbal subtests (such as *Similarities).* The 7 cluster solution created an additional math cluster containing only *Math Reasoning.*

Table 8

Three Cluster Solution

Cluster 1	*Cluster 2*	*Cluster 3*
Picture Arrangement	**Analysis/Synthesis (Deduction)**	Numerical Operations
Coding	Mystery Codes	**Math Reasoning (Quantitative Reasoning)**
Logical Steps (Deduction)	Classification	**Arithmetic (Quantitative Reasoning)**
Word Reasoning	**Picture Concepts (Induction)**	Applied Problems
Visual-Auditory Learning	Odd Item Out	Calculation
Planned Codes	Matrix Reasoning	
Digit Span	Story Completion	

Similarities	**Concept Formation (Induction)**
Symbolic Memory	Spatial Relations
Nonverbal Memory	Repeated Patterns
Verbal Spatial Relations	Seq. & Quant. Reasoning
	Pattern Reasoning
	Visual Coding
	Number Matrices
	Picture Completion
	Number Series

Table 9

Four Cluster Solution

Cluster 1	*Cluster 2*	*Cluster 3*
Applied Problems	Pic. Arrangment	Num. Operations
	Coding	**Math Reasoning (Quantitative Reasoning)**
Cluster 4	**Logical Steps (Deduction)**	**Arithmetic (Quantitative Reasoning)**
Analysis/Synthesis (Deduction)	Word Reasoning	Calculation
Mystery Codes	Vis-Aud Learning	
Classification	Planned Codes	
Picture Concepts (Induction)	Digit Span	
Odd Item Out	Similarities	
Matrix Reasoning	Symbolic Memory	
Story Completion	Nonverbal Memory	
Concept Formation (Induction)	Verb. Spat. Rel	
Spatial Relations		
Repeated Patterns		
Seq./Quant Reason.		
Pattern Reasoning		
Visual Coding		
Number Matrices		
Pic. Completion		
Number Series		

Table 10

Five Cluster Solution

Cluster 1
Picture Arrangment
Analysis/Synthesis (Deduction)
Logical Steps (Deduction)
Mystery Codes
Planned Codes
Similarities
Story Completion
Verbal Spatial Rel.
Picture Completion
Number Series

Cluster 4
Classification
Picture Concepts (Induction)
Odd Item Out
Matrix Reasoning
Concept Formation (Induction)
Spatial Relations
Repeated Patterns
Seq./Quant Reason
Pattern Reasoning
Visual Coding
Number Matrices

Cluster 2
Applied Problems

Cluster 5
Coding
Word Reasoning
Vis. Aud. Learning
Digit Span
Symbolic Memory
Nonverbal Memory

Cluster 3
Num. Operations
Math Reasoning (Quantitative Reasoning)
Arithmetic (Quantitative Reasoning)
Calculation

Table 11

Six Cluster Solution

Cluster 1
Picture Arrangment
Analysis/Synthesis (Deduction)
Logical Steps (Deduction)
Mystery Codes
Simlarities

Cluster 2
Classification
Picture Concepts (Induction)
Odd Item Out
Matrix Reasoning
Concept Formation

Cluster 3
Num. Operations
Math Reasoning (Quantitative Reasoning)
Arithmetic (Quantitative Reasoning)
Calculation

Story Completion
Picture Completion
Number Series

Cluster 4
Applied Problems

(Induction)
Spatial Relations
Repeated Patterns
Seq/Quant Reason
Pattern Reasoning
Visual Coding
Number Matrices

Cluster 5
Coding
Vis. Aud. Learning
Planned Codes
Symbolic Memory
Nonverbal Memory

Cluster 6
Word Reasoning
Verbal Spatial Rel.

Table 12

Seven Cluster Solution

Cluster 1
Pic.Arrangement

Analysis/Synthesis (Deduction)
Logical Steps (Deduction)
Mystery Codes
Similarities
Story Completion
Picture Completion
Number Series

Cluster 4
Applied Problems

Cluster 7
Classification
Picture Concepts (Induction)
Odd Item Out
Matrix Reasoning
Concept Formation (Induction)
Spatial Relations

Cluster 2
Math Reasoning (Quantitative Reasoning)

Cluster 5
Coding
Vis. Aud. Learning
Planned Codes
Digit Span
Symbolic Memory
Nonverbal Memory

Cluster 3
Num. Operations

Arithmetic (Quantitative Reasoning)
Calculation

Cluster 6
Word Reasoning
Verbal Spatial Rel.

Repeated Patterns
Seq/Quant Reason
Pattern Reasoning
Visual Coding
Number Matrics

Validity Check. The WJ-III Ach *Applied Problems* subtest, a task similar to the *Math Reasoning* subtest was included as a validity check to ascertain whether participants were being reasonably vigilant in their sorting. Since the task descriptions of these subtests were very similar, if participants were being vigilant, then these tasks should be located within the same clusters. As reported in Table 7, *Applied Problems* was located within the same cluster as *Math Reasoning* in the three cluster solution. However, these subtests did not appear in the same cluster in any other solution. It is interesting to note that other than in the three cluster solution, *Applied Problems* consistently was the only member in its cluster.

CHAPTER FIVE: DISCUSSION

Summary of Results

The purpose of this study was to assess whether judging subtest similarities based on task demand data lead to similar results when judging subtest similarity using ability data. It compared two methods of subtest classification, a confirmatory factor analysis using ability data, and a cluster analysis of similarity ratings of subtests' functional task demands. These two methods are salient in literature supporting practices in cross-battery assessment (e.g. Flanagan, Ortiz & Alfonso, 2007; Frisby & Parkin, 2007), where subtests from cognitive batteries are generally classified into broad ability categories through confirmatory factor analyses, while classified into narrow ability categories through functional task analysis processes.

Phase I – Ability Classification Through Factor Analysis. Results indicated that the originally planned three factor solution would not converge properly, because the correlations between the latent deductive reasoning factor with both the latent inductive and quantitative reasoning factors were estimated to be greater than 1, and error variance associated with the dedutive reasoning was estimated to be less than 0. However, reanalysis of a two factor solution specifying an inductive reasoning and a quantitative reasoning factor converged properly, and allowed for comparison with a one-factor model. An analysis of goodness-of-fit statistics indicated that both the one-factor and two-factor models provided excellent fits for the data. Nevertheless, the two-factor model provided a slightly better fit on many indexes and was retained as the best solution. This model placed *Arithmetic* and *Math Reasoning* on a quantitative reasoning factor, and *Picture Concepts* and *Concept Formation* on an inductive reasoning factor. These subtest pairs appear to measure different, albeit strongly related cognitive abilities.

Findings from this analysis confirm the narrow ability classifications suggested by other researchers. Flanagan and colleagues (2007) suggested that *Picture Concepts* measured inductive reasoning and *Arithmetic* reflected a measure of quantitative reasoning, classifications reported by Keith et. al. (2006). *Picture Concepts* loaded heavily on the same factor as *Concept Formation* a test confirmed in numerous studies as a measure of inductive reasoning (e.g. Flanagan & McGrew, 1998; Keith, Kranzler, & Flanagan, 2001; Phelps, McGrew, Knopik & Ford, 2005; Roid, 2003; Sanders et al., 2007; Woodcock, 1990). *Arithmetic* loaded strongly with the WIAT-II's *Math Reasoning* subtest, a measure used to assess math achievement skills which confirms the quantitative nature of its task demands.

Phase II - Classification of Functional Task Demands. Phase two of this study required practitioners to classify subtests by inferring the abilities a subtest may measure through a description of their functional task demands. These similarity judgements were subjected to series of non-hierarchical cluster analysis, where 3, 4, 5, 6 and 7 cluster solutions were specified. The relative locations of the three target pairs of subtests were tracked across solutions. Results indicated that it took a minimum of 5 clusters to place the subtests into their hypothesized pairs. The location of each subtest in the 3 target pairs varied across the number of solutions. For instance, the inductive reasoning pair, *Concept Formation* and *Picture Concepts* were placed within the same cluster in all solutions. The quantitative reasoning pair, *Arithmetic* and *Math Reasoning*, were placed within the same cluster in the initial 3 cluster solution, and stayed together in all solutions except the 7 cluster solution. The deductive reasoning pair, *Logical Steps* and *Analysis Synthesis*, were not placed in the same cluster until the five cluster solution. However, these two subtests continued to be placed together in the six and seven cluster solutions.

The validity check subtest pair, *Math Reasoning* and *Applied Problems,* were placed in the same cluster only in the three cluster solution. This could suggest that participants were not paying close attention to while they were sorting subtests, as these tasks' descriptions were very similar. However, it is interesting to note that after the three cluster solution, *Applied Problems* was consistently in its own cluster, not placed with any other subtest. It could be that participants considered this particular description to be unique when compared to the other descriptions, perhaps because it was selected from an achievement battery. However, *Math Reasoning* is also from an achievement battery.

It was possible to determine patterns among how tasks were sorted during the cluster analysis, though most cluster membership appeared highly variable. In general, clusters were most congruent in terms of the broad abilties they contained, but not in their narrow abilities. Specifically, inductive and deductive reasoning tasks were often included together. Visual-spatial tasks were also often included with reasoning tasks, results that have occurred in similar studies (Frisby & Parkin, 2007).

Comparison of Findings

Providing tentative support to the use of task analysis when classifying subtests into narrow ability categories, there appears to be some congruence between the results of the CFA and the cluster analysis. The 5 cluster solution, where the target subtests first grouped together, appears to discriminate between inductive, deductive and quantitative reasoning tasks. As mentioned earlier, there appears to also be significant variability within clusters. For instance, *Mystery Codes* (a test hypothesized to be a measure of inductive reasoning) was classified with *Logical Steps* and *Analysis/Synthesis*, this study's indicators of deductive reasoning. Similarly, *Visual Codes* (a hypothesized measure of deductive reasoning) was grouped with *Concept Formation* and *Picture Concepts,* measures of inductive reasoning. Thus, although the target tests clustered as expected in the 5 cluster solution, not every task included in the classification was sorted with tasks to which they were considered similar by cross-battery guidebooks.

It is possible that the variability in the clusters is related to participants clustering tasks more on their surface level features, rather than on the cognitive processes they measure. Evidence for this is equivocal. For instance, many of the math subtests sorted together, regardless of whether they were of an applied nature, or if they relied more on the ability to do rote math compuation, a finding that suggests participants were focusing more on surface level features. However, *Number Series,* and *Number Matrices* did not tend to sort with the other tests involving numbers. Since these tasks are considered measures of quantitative reasoning rather than knowledge abilities, it may suggest that participants were vigilent regarding the cognitive processes they require. Tasks related to processing speed and memory abilities tended to cluster together, likely because these abilities are easily discernable from reasoning and visual-processing types of tasks. One reason may be due to the fact that the difference between these tasks are related to the broad abilities they may measure, rather than narrow abilities.

Participants may have a more difficult time distinguishing between narrow abilities using task analysis than distinguishing between broad abilities. For instance, the cluster analysis results suggested that participants mixed inductive and deductive reasoning tasks. There appear to be two related reasons behind why inductive and deductive reasoning abilities were difficult to discriminate in the task analysis. First, these abilities tend to both be required in reasoning tasks. For example, in *Concept Formation*, considered a measure of inductive reasoning, the final step to solve a puzzle actually requires a deductive step to determine the right answer. Second, the results of the sorting task may be influenced by participants' level of knowledge of CHC theory. Participants who completed the sorting task successfully (e.g. had no missing data), had a mean rating of 3.16 when rating their knowledge of CHC theory, and rated their use of it in practice a 2.60. From these ratings, it appears that participants did not consider themselves especially knowledgeable of CHC theory. They may not use constructs such as fluid reasoning, induction or deduction in practice, and may not have been considering them when sorting the tasks. Kaufman (1994) highlighted that task analysis may be highly dependant on the analyzer's theoretical perspective. As a result, it is possible that clinicians very familiar with CHC theory may classify tasks more in line with CHC theory than clinicians who are less familiar with the theory. Also, since the validity check pair of subtests did not sort together consistently, it is also possible that participants were not highly vigalent during the sorting.

Implications of Findings

Results from these two investigations suggest partial support for a) the constructs of induction, deduction and quantitative reasoning narrow abilities, and b) similarity between two methods of subtest task classification. A number of issues limited interpretation of findings. First, the ability data used in this investigation could not allow for interpretation of a three-factor model to directly compare to a one-factor model, and a two-factor model had to be created. Second, task analysis data sorted as hypothesized when a five-cluster solution was specified, but the hypthesized sorting was not maintained in subsequent solutions with a larger number of clusters. Third, the validity check pair of subtests did not sort together consistently.

Findings from the confirmatory factor analysis support the interpretation of inductive reasoning and quantitative reasoning abilities as both distinct, but related abilities (Flanagan & McGrew, 1997; Flanagan, Ortiz & Alfonso, 2007 Flanagan, McGrew & Ortiz, 2000). This finding underscores the cross-battery assessment practice of interpreting tests at the broad ability level, unless the scores of the subtests are significantly discrepant from each other (Flanagan, Ortiz & Alfonso). Such practice acknowledges that narrow ability measures appear highly associated, though may be differently developed in some individuals. However, this interpretation must be made cautiously, as the three-factor solution, could not converge properly.

Similarities between the CFA and the cluster analysis suggests that task analysis may have potential as a method of narrow ability classification. It provides additional support to the expert consensus classification studies previously discussed (e.g. Flanagan, Ortiz, Alfonso, & Mascolo, 2001; McGrew, 1997). However, the cluster analysis used in this investigation is associated with significant limitations.

Limitations and Future Directions

Results from these analyses should be interpreted in light of significant limitations. The sample size used in the CFA reflects one limitation, and may be the reason the three-factor model was not positive definite (Brown, 2006; Marsh & Hau, 1999). Brown (2006) argues that small sample sizes are often more suspect to the influence of outliers, collinearities and poor approximations of normality. According to Brown (2006), and Marsh and Hau (1999), small sample sizes are at high risk for improper solutions when there are few indicators per latent variables (especially 2 indicators), a condition present in the current research. These authors recommend that CFA studies with small sample sizes should include numerous (e.g. 4, 5 or more) indicators per latent factor and that these indicators should all be strong (factor loadings higher than .6). Future research may use more subtests to account for narrow abilities. Relatedly, though this investigation'ssample's ability distribution did not department signifiantly from normality, greater representation of a full range of ability level may have also been helpful in avoiding a Heywood case, as it would have minimized a restriction of range.

Other limitations can be inferred from the results of Floyd et al. (2005), who demonstrated that score differences from batteries that presumably measuring the same construct are due to a number of factors. These researchers reported that in general, sampling characteristics, the date of a sample, and the scaling of a subtest (e.g. differences in standardized means and standard deviations) do not have a systematic influence on score differences, but battery floor and ceilings do appear associated with such differences. In this study, the WISC-IV is normed on youth between the ages of 6 and 16, while the WJ-III Cog is normed on individuals between the ages of 2 to 90+. As a result, the WJ-III Cog will have a lower floor and a higher ceiling. Though likely this effect was minimal on the results of this investigation, future research may wish to guard against this influence by using test batteries with more similar floors or ceilings.

The number of narrow abilities represented in this study further limit interpretation of its results. While fluid reasoning abilities were well represented in the tasks used in this investigation, it is likely that the target subtests also measure mulitple narrow abilities that were not adequately represented. For instance, researchers have debated whether *Arithmetic* measures a combination of math knowledge, quantitative reasoning and working memory skills (Keith et. al., 2006; Keith & Witta, 1997; Phelps et al., 2005). Similarly, success on *Math Reasoning* may also require significant quantitative knowledge skills. *Picture Concepts*, besides requiring inductive reasoning abilities, may also tap significant prior knowledge. Due to the limited focus of the present study, these analyses can not give a full understanding of the abilities these tests may measure. The variation within the cluster sorting also supports this assertion. For instance, *Number Series* and *Number Matrices,* subtests more associated with quantitative reasoning (Flanagan, Ortiz & Alfonso, 2007), were sorted with this study's deductive reasoning target tests.

There are also limitations associated with this investigation's method of task analysis. Discussed previously, participants in this investigation did not consider themselves experts on CHC theory. Differences between these results and previous task analyses reported in McGrew (1997) and Flanagan, Ortiz, Alfonoso and Mascolo (2001) may be associated with differences in participants' experience, or lack of experience, using CHC theory. Future research is needed to understand how participants' experience with CHC theory may alter how they interpret subtests. Future researchers may better investigate this influence by asking two groups of clincians, one group with signficant

CHC knowledge, and one without CHC level knowledge, to perform the sorting task and describe whether any differences exist.

The non-hierarchical analysis used in this investigation relied on random seeds as initial cluster solutions, which may also represent a limitation. Some scholars (e.g. Hair & Black, 2000) consider non-hierarchical analyses to be most useful when the investigator can provide initial estimates of cluster solutions, because solutions can be dependent on the initial seeds. However, it is interesting to note that the target subtests in this investigation generally cluster together even with random seeds were used, and maintained their cluster membership across mulitple solutions. Future research using similar methodology may wish to use the target subtests as centers for clusters to better determine which subtests would cluster with them.

Though the present study does indicate that there is a level of congruence between the two methods of narrow ability classification, it does not provide a measure of the degree of that congruence. Five clusters were necessary before the target subtests were sorted into the same clusters, but it is not possible to discern whether that is a high level of congruence, or only a moderate or low level. One way to gain greater understanding of the degree of congruence would be to follow procedures outlined in Frisby and Parkin (2007). These researchers used multidimensional scaling (MDS) methods to quantify a difference between different subtests' task demands on a three-dimensional space. Future research could follow their methods to quantify task demand differences, and then correlate those differences to differences between subtest ability correlations.

Conclusions

The presented study investigated two methods of subtest narrow ability classification, ability analyis and task analysis, while targeting fluid reasoning tasks. Both are prevalent in cross-battery assessment literature, used as methods to ensure adequate construct representation in the assessment process. Current results tentatively support the use of cross-battery assessment pillar III, (at least in terms of the narrow fluid reasoning abilities inductive and quantitative reasoning), because of overlap in the two classification processes.

Similar to those from Frisby and Parkin (2007), the present results provide an empirical method of understanding subtest task demands, a goal of researchers and clinicians alike (Carroll, 1976; Floyd, 2005; Kaufman, 1994). They also support the need for the integration of information-processing and psychometric research on cognitive tasks (e.g. Floyd, 2005). Floyd stresses that Carroll's coding scheme (1976) for cognitive tasks reflects an important stepping stone for investigating this area. To support that process, more research is need to validate the coding scheme and discern how it relates to CHC theory, but the methods outlined here could help with that process.

Though there may be a place for the rational process of individual clinicians when engaging in the task analysis process (e.g. Kaufman, 1994), empirical investigation of task demands may better validate what attributes of tasks are more interpretable than others. For instance, Kaufman has suggested that the both the length of an examinee response and the method of stimuli presentation are an important aspect of a task. But researchers do not yet know if they are *equally* important. Future research on task analysis may better answer these questions.

APPENDIX A

Subtest Descriptions

Induction

Concept Formation	WJ-III Cog

19) The examinee views a page containing rows of shapes that differ according to a combination of size (big, small), color (red, yellow), and shape (circles, squares, triangles). To answer, the examinee must articulate the rule that explains why some shapes in a row are enclosed in a box while others in the row are not.

Picture Concepts	WISC-IV

10) The examinee views two or three rows of pictures of concrete objects on a page. To answer, the examinee points to or names one picture from each row that belong in the same category.

Pattern Reasoning	KABC-II

24) The examinee views a page showing a pattern of geometric, abstract, or concrete figures, where an object in each pattern is missing. To answer, the examinee must determine the pattern and use that information to point to, or name a response from a row of choices the best completes the pattern. Examinees receive bonus points for fast answers, though there is no time limit.

Story Completion	KABC-II

18) The examinee views a row of pictures that tell an incomplete story (one or more pictures are missing in the story). To answer, the examinee must determine the pattern demonstrated in the story and complete the story by selecting a picture(s) from a number of cards presented to them by the examiner. Items are timed.

Mystery Codes	KAIT

7) The examinee views a page that demonstrates a code associating symbols with different characteristics of an abstract or concrete figure. To answer, the examinee must determine the code in order to apply it to a new picture, and point to, name or circle the correct code. Examinees can use paper and pencil to help them solve the problems. Items are timed.

Classification	Leiter-R

9) The examinee views a row of pictures depicting concrete or abstract figures. The examinee is presented with a number of cards that have similar figures on them. To answer, the examinee must categorize these cards by placing them in front of the figure with which they share an essential similarity.

Repeated Patterns Leiter-R

21) The examinee views a page with a number of rows that depict a pattern of abstract or concrete figures, with part of the pattern missing. The examinee is presented with a number of cards that have similar figures on them. The examinee must determine the pattern in the page and use that information to select the missing parts of the pattern from the cards. To answer, the examinee places the correct cards in the missing part of the pattern.

Matrix Reasoning WISC-IV

15) The examinee views a page with a matrix of abstract figures with one section missing. Five possible responses are listed on the bottom of the page. The examinee must determine the relationship between the figures in the matrix and then use that relationship to decide which of the five responses best fits in the matrix. To answer, the examinee points at or names their choice from the five responses.

Deduction

Analysis/Synthesis WJ-III Cog

4) The examinee views pages containing squares of different colors. The examiner teachers the examinee a variety of equations about relationships between colors (e.g., a blue square is the same as a yellow square with a red square). The examinee is shown additional pages that show groups of colored squares arranged in equations, with a missing, blank square. To answer, the examinee names the color that belongs in the blank square, according to the color equations learned previously.

Logical Steps KAIT

5) The examiner reads a set of rules or premises about the relationships between different characters in a diagram to the examinee while the examinee views the diagram. To answer, the examinee uses these premises to verbally provide responses to questions about characters' relative placement within that diagram. Examinees may use scratch paper to solve the problems. Items are timed.

Visual Coding Leiter-R

28) The examinee views an easel page containing abstract and concrete stimuli. The examinee is taught a number of equations (e.g. a spoon goes with a star, or a pencil is the same as a circle and a triangle). Examinees must use these equations to answer items

printed at the bottom of the page. Potential responses are printed on cards. To answer, the examinee must select the correct card and place it next to the item it answers.

Quantitative Reasoning

Number Matrices WJ-III

30) The examinee views a page that contains a two-dimensional matrix of numbers with a one number missing. The examinee looks for a relationship rule between the numbers, and then must identify the number that satisfies the rule to fill the missing section. To answer, the examinee verbally states the missing number.

Number Series WJ-III

32) The examinee views a page that shows a row of numbers with a number missing. The examinee must determine a pattern within the series of numbers. To answer, the examinee verbally states the missing number.

Arithmetic WISC-IV

17) The examinee hears a math word problem read by the examiner. To answer, the examinee verbally states their response to the problem, without the use of scratch paper. Each item is timed.

Sequential & Quantitative Reasoning DAS-II

22) The examinee sees three boxes containing pairs of numbers, however the last box has one number missing. The examinee determines the relationship between the pairs of numbers in the first two boxes and uses that information to determine the missing number in the third box. To answer, the examinee verbally states the missing number.

Mathematics Reasoning WIAT-II (VALIDITY)

14) The examinee views a page depicting a math story problem. The examinee must figure out the type of math operations necessary to solve the story problem. The examinee must then apply those operations to the information in the story. To answer, the examinee must verbally state their answer. The examinee may use scratch paper if they wish.

Applied Problems WJ-III Achievement (VALIDITY)

25) The examinee views a page showing a math story problem. The examinee must determine the type of math operations required by the story problem, and then apply those operations to the information in the story. To answer, the examinee must verbally state their response. The examinee may use scratch paper if they wish.

Distractors

Nonverbal Memory RIAS Gv Visual Memory

26) The examiner shows the examinee a target abstract or concrete figure for 5 seconds. To answer, the examinee points to the figure they saw when it is embedded within an array of similar, distractor figures.

Symbolic Memory UNIT Gv Visual Memory

23) The examiner presents the examinee with tiles of symbolic figures that vary in color (green & black), sex (man & woman) and age (boy/man & girl/woman). The examiner creates a sequence of these tiles. To answer, the examinee must recreate the sequence from memory.

Odd-Item-Out RIAS Gv Visualization (Vz)

13) The examinee views a page containing a number of abstract figures. One figure is different from the other items on the page. To answer, the examinee must point to the picture that does not belong with the others.

Picture Arrangement WISC-III Gv Visualization

1) The examiner presents the examinee with a number of cards. Each card depicts a part of a story. To answer, the examinee must arrange the cards in an order that tells a logical story. Each item is timed.

Picture Completion WISC-IV Gv Flexibility of Closure

31) The examinee views a page containing a common picture or scene. The examinee must determine what part of the picture or scene is missing. To answer the examinee must point to or name the missing part of the picture. Each item is timed.

Spatial Relations WJ-III Cog Gv Spatial Relations

20) The examinee views a page that shows a row of shape fragments that are part of a whole and a number of distractor fragments. The examinee must determine which fragments are used to create the whole. To answer, the examinee must verbally state or point to the correct fragments.

Visual-Auditory Learning WJ-III Cog Gsm Associative Memory

8) The examinee views a page with a number of simple figures printed on it. The examiner teaches the examinee a code by pointing to a figure and saying its

corresponding name. Next, the examiner shows the examinee a sequence of the figures. To answer, the examinee must remember the code to recite the names of the pictures.

Digit/Symbol Coding WISC-IV Gs Rate-of-Test-Taking

2) The examinee views a page that presents a code associating a number with a simple symbol. The examinee must use the code to draw the symbols in boxes associated with each number. To answer, the examinee must copy as many symbols as possible under a time limit, and in a sequential manner.

Planned Codes CAS Gs Rate-of-Test-Taking

11) The examinee views a page that presents a code associating a letter with a answer that consists of X's and O's. The examinee is told to use the code to fill in boxes associated with each letter. To answer, the examinne must copy as many codes as possible under a time limit. At the end of the time limit, the examinee explains the strategy they used to complete the task.

Verbal-Spatial Relations CAS Gc Listening Ability

29) The examinee views a page the shows six similar pictures consisting of either abstract or concrete figures. The examinee also listens to a question printed at the bottom of a page that is read by the examiner. The question refers to spatial relationships depicted in the pictures. To answer, the examinee must point at, or verbally state the picture that best answers the question.

Digit Span WISC-IV Gsm Memory Span

12) The examinee hears a list of single digit numbers, read by the examiner. To answer, the examinee must repeat the string of digits in either a forward or backward order.

Calculation WJ-III Achievement Gq Mathematical Achievement

27) The examinee is presented with a worksheet that includes progressively more challenging math problems. To answer, the examinee writes the answer to as many math problems as he or she can.

Numerical Operations WIAT Gq Mathematical Achievement

3) The examinee solves increasingly difficult math problems while completing a worksheet. To answer, the examinee writes the answer to as many math problems as possible.

Similarities WISC-IV Gc

16) The examinee hears two objects or concepts read by the examiner. The examinee must determine how the two objects or concepts are alike. To answer, the examinee responds verbally with a brief answer.

Word Reasoning	WISC-IV	Gc

6) The examinee listens to a series of clues about an unknown object or concept read by the examiner. The examinee uses the clues to guess the object or concept. To answer, the examinee verbally states the response.

APPENDIX B

Data Sheets

Instructions:

1) Fill out the demographic form on page 2 of this packet.

2) You have received a packet of 32 descriptions of subtests from common cognitive test batteries. Please sort these descriptions into piles based on similarities you see in the cognitive abilities these tests measure – do NOT sort them based on the cognitive battery they come from. You can sort the descriptions into **as many categories as you wish**. However, you must make *at least 3* categories. Also, each category must contain *at least* 2 cards.

3) After you have sorted all 32 descriptions, please record your responses on the data sheet provided on page 3.

   Record your responses by writing down the number of the subtest descriptions in each category next to each other on this sheet.

   For example:

Category 1	Category 2	Category 3	Category X
3	8	2	9
32	14	12	1
19	17		31
	28		
	4		
	18		
	29		

Remember to make *at least 3* categories. Each category must contain *at least 2* descriptions.

Demographic Information

Date: __________ Age: __________ Sex: __________

Please Check All That Apply

______________ School Psychology Graduate Student
______________ Credentialed School Psychological Examiner
______________ Credentialed School Psychologist
______________ Licensed Psychologist
______________ Educational Diagnostician
______________ Other

District Employed (If Applicable) ____
University Enrolled (If Applicable) ____

Number of Years as a Paid Practitioner

Please indicate the cognitive (intelligence) scales on which you have had prior experience administering (Check all that apply):

______________ Bayley Scales of ID
______________ Bayley Scales of ID-II
______________ CAS
______________ CTONI
______________ DAS
______________ DAS-II
______________ DTLA-3
______________ DTLA-4
______________ KABC
______________ KABC-II
______________ KAIT
______________ K-BIT
______________ Leiter IPS-R

__________ RIAS (Reynolds)
__________ SIT-R (Slosson)
__________ Stanford-Binet IV
__________ Stanford-Binet V
__________ UNIT
__________ WAIS-III
__________ WISC-III
__________ WISC-IV
__________ Woodcock Johnson R (Cog)
__________ Woodcock Johnson-III (Cog)
__________ WPPSI-R
__________ WPPSI-3

Please read the following statements are circle the number that most corresponds with your atttude.

I am knowledgeable about the Cattell-Horn-Carroll (CHC) theory of cognitive abilities.

1	2	3	4	5
Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

I use the Cattell-Horn-Carroll (CHC) theory of cognitive abilities when interpreting test results.

1	2	3	4	5
Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

Data Record Sheet

Record your responses by writing down the number of the subtest descriptions in each category that you create next to each other on this sheet.

For example:

Category 1	Category 2	Category 3	Category X
3	8	2	9
32	14	12	1
19	17		31
	28		
	4		
	18		
	29		

Please record your responses below. Make as many categories as you like, but remember to make *at least 3* categories. Each category must contain *at least 2* descriptions. Remember, do NOT sort them based on the cognitive batteries they may come from.

Informed Consent

This form requests your consent to participate in a research study investigating task demands similarities of cognitive subtests used in school psychology assessment practice. The project is under the direction of Craig Frisby, Ph.D, and has been approved by the University of Missouri Institutional Review Board.

Project description: Your participation involves a single meeting with a graduate student. To participate, you will provide general demographic information, and sort a stack of cards that contain descriptions of cognitive subtests into a number of categories and record your responses on record sheets. This task should require no more than 20 to 30 minutes. There will be approximately 50 participants involved in this study. Individuals eligible to participate include individuals familiar with the administration and interpretation of common cognitive batteries.

Potential Benefits and Concerns: There are neither any material benefits nor risks anticipated greater than those encountered in daily life. Results of the project may be useful to participants as knowledge relevant to school psychology practice. If you decide to participate, you will be entered into a drawing for a $20 Starbucks Giftcard.

Confidentiality: All information regarding your responses will be kept confidential according to legal and ethical guidelines. Your name will not appear on any research protocols, but will be replaced by a code number. No one that is not connected with this project will have access to these files. In addition, any information gathered will not be shared with any outside person or agency unless for any reason you give us permission to do so. After the project is complete, all data will be aggregated, making it impossible to identify one person's test data. No names of participants are used in any published material that may result from the study.

Participation is Voluntary: Your participation is entirely voluntary. You are free not to answer any question you do not choose to answer. You can freely withdraw from the project at any time without negative consequences. After the completion of the study, all data pertaining to your participation will be kept for three years and then destroyed.

Questions? Please call Dr. Craig Frisby at 573-884-2561 or Jason Parkin at (425)890-2404 with any questions or concerns. If you have questions about your rights as a research project participant, you may contact the MU Institutional Review Board at 573-882-9585.

Please sign below to indicate that you have read and understand the above information.

Signed: ____ Date: __________________

Verbal Instructions Script

I am conducting a research study on the similarities of the task demands of cognitive subtests used in school psychology assessment practice. The project is sponsored by the University of Missouri-Columbia under the direction of Craig Frisby, Ph.D. The project has been approved by the University of Missouri Institutional Review Board.

If you choose to participate at this time I'll ask you to do the following: first, you will complete a brief sheet requesting general demographic information. Next, you will be given a stack of cards that contain descriptions of cognitive subtests. I'd like you to sort the cards into similar categories and record your responses on the answer sheets provided. Participation should not require more than half an hour.

There are neither any material benefits nor risks anticipated greater than those encountered in daily life. We will be happy to share results of the project after it is complete. Also, a thank you for participating, we'll have a quick lottery for a $20 Starbucks gift certificate afterwards.

All information regarding your answers will be kept confidential. Your name will not appear on any research protocols study, but will be replaced by a code number. All protocols will be kept in locked files in my office at the University of Missouri. No one that is not connected with this project will have access to these files. In addition, any information gathered will not be shared with any outside person or agency unless for any reason you give us permission to do so. After the project is complete, all data will be aggregated, making it impossible to identify one person's test data. No names of participants are used in any published material that may result from the study.

Your participation is entirely voluntary. You are free not to participate. If you participate, you are free not to answer any question you do not choose to answer. You can freely withdraw from the project at any time without negative consequence. After completion of the study, all data pertaining to your participation will be kept for three years and then destroyed. Participation or non-participation will not affect (the course grades of university students who volunteer in this research, or) evaluations of employers.

Should you have questions, please call Dr. Craig Frisby at 573-884-2561 with any questions or concerns. If you have questions about your rights as a research project participant, you may contact the MU Institutional Review Board at 573-882-9585.

The consent forms I will pass out includes the information I have just shared with you.

APPENDIX C

Are you between the ages of 11 and 16? Earn a $5 Gift Certificate to McDonalds!

Q: What is this for?

A: *This is a study that looks at how kids can do different kinds of puzzles and solve problems.*

Q: What will I do?

A: *You will take six brief tests with a student from the University of Missouri. The tests are puzzles that you will figure out by looking at shapes, or solving short math problems. The problems start out easy, but can get harder. Try your best!*

Q: How long will it take?

A: *It will take about 45 minutes to do all six types of puzzles.*

Q: How many kids will be in the study?

A: *About 60 kids will be in this study.*

Q: Are there any risks?

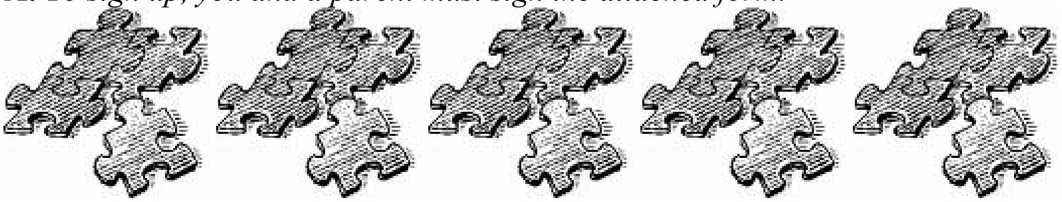
A: *Your participation should not cause any more risk than you usually experience at school. Some kids may feel frustrated when completing the puzzles. They start out easy, but can get hard.*

Q: What will I get?

A: *Many kids think these puzzles are fun to do! Also, as a "thank you" for your participation, you will receive a $5 gift certificate to McDonalds!*

Q: How do I sign up?

A: *To sign up, you and a parent must sign the attached form.*



Dear Sir or Madam:

The purpose of this form is to request your permission to include your student in a research project investigating the interpretation of intelligence tests. The project is under the direction of Craig Frisby, Ph.D, and has been approved by the University of Missouri Institutional Review Board.

Project Description. Your student's participation will involve meeting with a graduate student in school psychology from the University of Missouri. The graduate student will administer six brief tests that assess your student's ability to reason and solve novel problems. The graduate student will also collect general demographic data from your student. This meeting should take approximately 45 minutes.

Risks and Benefits. There are no risks associated with this project that are greater than those that occur in everyday life. Your student will be taking tests that begin with an easy difficulty level, though may become difficult for most youth. Most youth will not answer every question correctly, and though they are not told whether they are right or wrong, some individuals may experience mild frustration during this process. As a token of appreciation, your student will receive a samll gift certificate to McDonalds. Results of this study will inform school psychologists and other educational professionals about the similarities and differences between common tests they use in practice, and may aid in a more accurate interpretation of results from these tests.

Confidentiality: All information regarding your students' responses will be kept confidential according to legal and ethical guidelines. Their name will not appear on any research protocols, but will be replaced by a code number. All protocols will be kept in locked files in Jason Parkin's office at the University of Missouri. It will be destroyed after 3 years. No one that is not connected with this project will have access to these files. In addition, any information gathered will not be shared with any outside person or agency for any reason. After the project is complete, all data will be aggregated, making it impossible to identify one person's test data. No names of participants will be used in any published material that may result from the study.

Participation is Voluntary: Your student's participation is entirely voluntary. You may freely withdraw them from the project at any time without negative consequences. After the completion of the study, all data pertaining to their participation will be kept for three years and then destroyed.

Questions? Please call Dr. Craig Frisby at 573-884-2561 with any questions or concerns. If you have questions about your rights as a research project participant, you may contact the MU Institutional Review Board at 573-882-9585.

If your student is interested in participating, please sign this form. Keep the above information for your records.

Parent Name (Please Print)

Signature

Date

Student Name (Please Print)

Student Signature

Date

REFERENCES

Atkinson R.C., & Shiffrin, R.M. (1968). Human memory: A proposed system and its control processes. In K.W. Spence & J.T. Spence (Eds.), *The Psychology of learning and motivation* (pp. 89-195). New York: Academic Press.

Brown, T.A. (2006). Confirmatory Factor Analysis for Applied Research. New York: Guilford Press.

Bryant, F.B., & Yarnold, P.R. (1995). Principal-components analysis and exploratory and confirmatory factor analysis. In L.G. Grimm & P. R.Yarnold's (Eds.) *Reading and Understanding Multivariate Statistics* (pp. 99-136). Washington, D.C: APA.

Buckland, J.A., McGee, R.L. & Ehrler, D.J. (2001). An investigation of *Gf-Gc* theory in the older adult population: Joint factor analysis of the Woodcock-Johnson-Revised and the Detroit Tests of Learning Aptitude-Adult. *Psychological Reports, 88,* 1161-1170.

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.* Cambridge Cambridge, MA: University Press.

Carroll, J.B. (1997). The three-stratum theory of cognitive abilities. In D.P. Flanagan, J.L. Genshaft & P.L. Harrison's (Eds). *Contemporary Intellectual Assessment: Theories, tests and issues*. (pp. 122-130). New York: Guilford Press.

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

Christou, C., & Papageorgiou, E. (2007). A framework of mathematics inductive reasoning. *Learning and Instruction, 17,* 55-66.

Chronbach, L.J. & Snow, R.E. (1997). *Aptitudes and Instructional Methods.* New York: Irvington.

Cohen, R.J. & Swerdlik, M.E. (2002). *Psychological Testing and Assessment: An Introduction to Test and Measurement, Fifth Edition.* McGraw Hill. Boston, MA.

DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods, 2*, 292-307.

Elliot, C.D. (1990). *Differential Ability Scales.* San Antonio, TX: The Psychological Corporation.

Elliot, C.D. (2007). *Differential Ability Scales, Second Edition*. San Antonio, TX: The Psychological Corporation.

Esters, I.G., Ittenbach, R.F., & Han, K. (1997). Today's IQ tests: Are they really better than their historical predecessors? *School Psychology Review, 26,* 211-224.

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.

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

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

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.

Flanagan, D.P., & McGrew, K.S. (1998). Interpreting intelligence tests from contemporary *Gf-Gc* theory: Joint confirmatory factor analysis of the WJ-R and KAIT in a non-white sample. *Journal of School Psychology, 36,* 151-182.

Flanagan, D.P., McGrew, K.S. & Ortiz, S.O. (2000). *The Wechsler Intelligence Scales and Gf-Gc Theory: A Contemporary Approach to Interpretation.* Boston, MA: Allyn & Bacon.

Flanagan, D.P., Ortiz, S.O., Alfonso, V.C., & Mascolo, J.T. (2001). *The Achievement Test Desk Reference: Comprehensive Assessment and Learning Disabilities.* Boston, MA: Allyn & Bacon.

Flanagan, D.P., Ortiz, S.O., & Alfonso, V.C. (2007). *Essentials of Cross-Battery Assessment, Second Edition*. Hoboken, NJ: John Wiley & Sons, Inc.

Floyd, R.G. (2005). Information-processing approaches to interpretation of contemporary intellectual assessment instruments. In D.P. Flanagan & P.L. Harrison's (*Eds.*) *Contemporary Intellectual Assessment, Second Edition (pp. 203-233).* New York: The Guilford Press.

Floyd, R.G., Berman, R., McCormack, A.C., Anderson, J.L., & Hargrove-Owen, G.L. (2005). Are Cattell-Horn-Carroll Broad Ability Composite Scores Exchangeable Across Batteries? *School Psychology Review, 34,* 329-357.

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.

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

Fiorello, C.A., & Primerano, D. (2005). Research into practice: Cattell-Horn-Carroll cognitive assessment in practice: Eligibility and program development issues. *Psychology in the Schools, 42*, 525-536.

French, J.W., Ekstrom, R.B., & Price, L.A. (1963). *Kit of reference tests for cognitive factors.* Princeton, N.J.: Educational Testing Services.

Frisby, C. L., & Parkin, J. R. (2007). Identifying similarities in cognitive subtest functional requirements: An empirical approach. *Journal of School Psychology, 45,* 385-400.

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. *The Journal of Special Education, 40,* 103-114.

Glutting, J.J., Watkins, M.W., & Youngstrom, E.A. (2003). Multifactored and cross-battery ability assessments: Are they worth the effort? In C.R. Reynolds & R.W.

Kamphaus (Eds.). *Handbook of Psychological and Educational Assessment of Children: Intelligence, Aptitude, and Achievement, Second Edition*. (pp.343-374). New York: The Guilford Press.

Goldman, S.R. & Pellegrino, J.W. (1984). Deductions about induction: Analyses of developmental and individual differences. In R.J. Sternberg *(Eds.) Advances in the Psychology of Human Intelligence, Vol 2.* (pp.149-198). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hair, J.R. & Black, W.C. (2000). Cluster analysis. In L.G. Grimm & P.R. Yarnold's *Reading and Understanding More Multivariate Statistics.* Washington, D.C.: APA.

Hammell, D.D. (1998). *Detroit Tests of Learning Aptitude- Fourth Edition*. Austin, TX: Pro-Ed.

Hammell, D. & Bryant, B. (1991). *Detroit Tests of Learning Aptitude – Adult.* Austin, TX: Pro-Ed.

Horn, J.L. (1998). A basis for research on age differences in cognitive capabilities. In J.J. McArdle & R.W. Woodcock (Eds.), *Human Cognitive Abilities in Theory and Practice* (pp. 57-87). Mahwah, NJ: Erlbaum.

Horn, J.L. & Noll, J. (1997). Human cognitive capabilities: Gf-Gc Theory. In D.P. Flanagan, J.L. Genshaft & P.L. Harrison's (Eds). *Contemporary Intellectual Assessment: Theories, tests and issues*. (pp. 53-91). New York: Guilford Press.

Hu, L. & Bentler, P.M. (1999). Cut-off criteria for fit indexes incovariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1-55.

Jensen, A.R. (1992). Commentary: Vehicles of *g*. *Psychological Science, 3,* 275-278.

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

Johnson-Laird, P.N. (1999). Deductive reasoning. *Annual Review of Psychology, 50,* 109-135.

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.

Kamphaus, R.W., Petoskey, M.D. & Morgan, A.W. (1997). A history of intelligence test interpretation. In D.P. Flanagan, J.L. Genshaft & P.L. Harrison's (Eds). *Contemporary Intellectual Assessment: Theories, tests and issues*. (pp. 32-47). New York: Guilford Press.

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

Kaufman, A.S., Lichtenberger, E.O., Fletcher-Janzen, E., & Kaufman, N.L. (2005). *Essentials of KABC-II Assessment*. Hoboken, NJ: John Wiley & Sons, Inc.

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

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

Keith, T.Z. (2005). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In D.P. Flanagan and P.L. Harrison (Eds.) *Contemporary Intellectual Assessment: Theories, tests and issues, Second Edition*. (pp. 581-624). New York: Guilford Press.

Keith, T.Z., Fine, J.G., Taub, G.E., Reynolds, M.R., & Kranzler, J.H. (2006). Higher order, multisample, confirmatory factor analysis with the Wechsler Intelligence Scale for Children – Fourth Edition: What does it measure? *School Psychology Review, 35,* 108-127.

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

Keith, T. Z., & Witta, L. (1997). Hierarchical and cross-age confirmatory factor analysis of the WISC-III: What does it measure? School Psychology Quarterly, 12, 89-107.

Klem, L. (2000). Structural equation modeling. In L.G. Grimm & and P. R. Yarnold (Eds.) Reading and Understanding More Multivariate Statistics. (pp. 227-260). Washington D.C.: APA.

Kline, R. B. (2004). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press.

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

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

Marsh, H.W. & Hau, K.T. (1999). Confirmatory factor analysis: Strategies for small sample sizes. In R.H. Hoyle's (Ed). *Satistical Strategies for Small Sample Research.* (pp. 251-284). Sage Publications: Thousand Oaks, CA.

McGhee, R. (1993). Fluid and crystallized intelligence: Confirmatory factor analyses of the Differential Abilities Scale, Detroit Tests of Learning Aptitude-3, and Woodcock-Johnson Psycho-Educational Battery-Revised. In R.S. McCallum & B.A. Bracken (Eds.), *Woodcock-Johnson Psycho-Educational Battery-Revised* (pp. 20-38). Knoxville, TN: Psychoeducational Corp.

McGrew, K.S. (1997). Analysis of the major intelligence batteries according to a proposed comprehensive Gf-Gc Framework. In D.P. Flanagan, J.L. Genshaft & P.L. Harrison's (Eds). *Contemporary Intellectual Assessment: Theories, tests and issues*. (pp. 151-179). New York: Guilford Press.

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

McGrew, K.S. & Flanagan, D.P. (1997). Beyond g: The impact of Gf-Gc specific cognitive abilities research on the future use and interpretation of intelligence tests in the schools. *School Psychology Review, 26,* 189-210.

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

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

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. *Journal of Special Education, 25,* 504-526.

McDermott, P.A. & Glutting, J.J. (1997). Informing stylistic learning behavior, disposition, and achievement through ability subtests—or, more illusions of meaning? *School Psychology Review, 26,* 163-175.

McDermott, Paul A; Fantuzzo, John W; Glutting, Joseph J. (1990). Just say no to subtest analysis: A critique on Wechsler theory and practice. *Journal of Psychoeducational Assessment,* 8, 290-302.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performance as a scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

Muthen, L. K., & Muthen, B. O. (2007). *Mplus version 5*. [Computer software]. Los Angeles: Muthen and Muthen.

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

Oh, H. J., 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. *The Journal of Special Education, 38*, 159–173.

Pellegrino, J.W. & Glaser, R. (1982). Analyzing aptitudes for learning: Inductive reasoning. In R. Glaser (*Eds.) Advances in Instructional Psychology, Vol 2.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Pfeiffer, S.I., Reddy, L.A., Kletzel, J.E., Schmeizer, E.R. & Boyer, L.M. (2000). The practitioner's view of IQ testing and profile analysis. *School Psychology Quarterly, 15,* 376-385.

Phelps, L., McGrew, K.S., Knopik, S.N., & Ford, L. (2005). The general (*g*), broad and narrow CHC stratum characteristics of the WJ III and WISC-III tests: A confirmatory cross-battery investigation. *School Psychology Quarterly, 20,* 66-88.

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

Reynolds, M, R., Keith, T.Z., Fine, J.G., Fisher, M.E. & Low, J.A. (2007). Confirmatory factor structure of the Kaufman Assessment Battery for Children – Second Edition: Consistency with Cattell-Horn-Carroll Theory. *School Psychology Quarterly, 22,* 511-539.

Rips, L.J. (1984). Reasoning as a central intellective ability. In R.J. Sternberg *(Eds.) Advances in the Psychology of Human Intelligence, Vol 2.* (pp.105-147). Hillsdale, NJ: Lawrence Erlbaum Associates.

Roid, G.H. (2003). *Stanford Binet Intelligence Scales, Fifth Edition.* Itasca, IL: Riverside Publishing.

Roid, G.H., & Miller, L.J. (1995, 1997). *Leiter International Performance Scale-Revised.* Wood Dale, IL: Stoelting, Co.

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

Schrank, F.A. & Flanagan, D.P. (2003). *WJ III Clinical Use and Interpretation.* San Diego, CA: Academic Press.

Spearman, C.E. (1904). “General intelligence,” objectively determined and measured. *American Journal of Psychiatry, 15,* 201-293.

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

SPSS, Inc. (2002). SPSS 11.5 user's guide. Chicago, IL: SPSS, Inc.

Sternberg, R. (1981). Testing and cognitive psychology. *American Psychologist, 36,* 1181-1189.

Sternberg, R.J. & Gardner, M.K. (1984). Unities in inductive reasoning. *Journal of Experimental Psychology: General, 112,* 80-116.

Stalans, L.J. (1995). Multidimensional scaling. In L.G. Grimm & P.R. Yarnold (Eds.) *Reading and Understanding Multivariate Statistics.* Washington, D.C.: American Psychological Association.

Tabachnick, B.G. & Fidell, L.S. (2007). *Using Multivariate Statistics* (5th ed.). Boston: Pearson.

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

Thorndike, R.L., & Hagen, E. *Cognitive Abilities Test.* Boston, MA: Houghton Mifflin.

Tusing, M.E. & Ford, L. (2004). Examining preschool cognitive abilities using a CHC framework. *International Journal of Testing, 4,* 91-114.

Watkins, M.W., Greenawalt, C.G., & Marcell, C.M. (2002). Factor structure of the Wechsler Intelligence Scale for Children-Third Edition among gifted students. Educational and Psychological Measurement, 62, 164-172.

Watkins, M.W., Glutting, J.J., & Lei, Pui-Wa. (2007). Validity of the Full-Scale IQ when there is significant variability among WISC-III and WISC-IV factor scores. *Applied Neuropsychology, 14,* 13-20.

Watkins, M.W., Glutting, J.J., & Youngstrom, E. A. (2005). Issues in subtest profile analysis. In D.P. Flanagan & P.L. Harrison's (*Eds.*) *Contemporary Intellectual Assessment, Second Edition (*pp 251-268*).* New York: The Guilford Press.

Watkins, M. W., Glutting, J. J., & Youngstrom, E. A. (2003). Cross-battery cognitive assessment: Still concerned. *NASP Communique, 31*, 42-44.

Watkins, M. W., Youngstrom, E. A., & Glutting, J. J. (2002). Some cautions concerning cross-battery assessment. *NASP Communique, 30*, 16-20.

Wechsler, D. (1974). *Wechsler Intelligence Scale for Children – Revised.* San Antonion, TX: The Psychological Corporation.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children – Third Edition.* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2001). *Wechsler Individual Achievement Test – Second Edition*. San Antonio, TX: The Psychological Corporation.

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

Wechsler, D. (2003b). *Wechsler Inlelligence Scale for Children-Fourth Edition: Technical and interpretive manual.* San Anionio. TX: Psychological Corporation.

Woodcock, R.W. (1998). Extending Gf-Gc theory into practice. In J.J. McArdle & R.W. Woodcock's (Eds.) *Human Cognitive Abilities in Theory and Practice*. (pp. 137-156).

Woodcock, R.W., & Johnson, M.B. (1989). *Woodcock-Johnson Psycho-Educational Battery – Revised.* Allen, TX: DLM Teaching Resources.

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

Youngstrom, E.A., Kogos, J.L., & Glutting, J.J. (1998). Incremental efficacy of Differential Ability Scales factor scores in predicting individual achievement criteria. *School Psychology Quarterly, 14,* 26-39.

VITA

Jason Richard Parkin was born in Downers Grove, Illinois on December 8th, 1979, the son of Richard Bradford Parkin and Katherine Grinde Parkin. After graduating from The International School, Bellevue Washington in 1998, he studied psychology at the University of Puget Sound, and graduated with a Bachelors of Arts in 2002. In August of 2003, he entered the doctoral program in school psychology at the University of Missouri. He received his Masters of Arts degree in May of 2007, and completed an APA accredited internship with the Lewisville Independent School District, Lewisville, TX in July of 2009.