Longitudinal Factor Structure of the Wechsler Intelligence Scale for Children-Fourth Edition in a Referred Sample

by

Lindsay Patricia Richerson

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved June 2012 by the
Graduate Supervisory Committee:

Marley Watkins, Co-Chair
John Balles, Co-Chair
Christa Lynch

ARIZONA STATE UNIVERSITY

August 2012

ABSTRACT

Standardized intelligence tests are some of the most widely used tests by psychologists. Of these, clinicians most frequently use the Wechsler scales of intelligence. The most recent version of this test for children is the Wechsler Intelligence Scale for Children - Fourth Edition (WISC-IV); given the multiple test revisions that have occurred with the WISC, it is essential to address evidence regarding the structural validity of the test; specifically, that the internal structure of the test corresponds with the structure of the theoretical construct being measured. The current study is the first to investigate the factor structure of the WISC-IV across time for the same individuals. Factorial invariance of the WISC-IV was investigated using a group of 352 students eligible for psychoeducational evaluations tested, on average, 2.8 years apart. One research question was addressed: Does the structure of the WISC-IV remain invariant for the same individuals across time? Using structural equation modeling methods for a four-factor oblique model of the WISC-IV, this study found invariance at the configural and weak levels and partial invariance at the strong and strict levels. This indicated that the overall factor structure remained the same at test and retest with equal precision of the factor loadings at both time points. Three subtest intercepts (BD, CD, and SI) were not equivalent across test and retest; additionally, four subtest error variances (BD, CD, SI, and SS) were not equivalent across test and retest. These results indicate that the WISC-IV measures the same constructs equally well across time, and differences in an individual's cognitive profile can be safely interpreted as reflecting change in the underlying construct across time rather than variations in the test itself. This allows clinicians to be more confident in interpretation of changes in the overall cognitive profile of individual’s across time. However, this study’s results did not indicate that an individual’s test scores should be compared across time. Overall, it was concluded that there is partial measurement invariance of the WISC-IV across time, with invariance of all factor loadings, invariance of all but three intercepts, and invariance of all but four item error variances.

DEDICATION

I dedicate this work to my amazing family and friends that have loved and supported me throughout this process. Without your love, prayers, and commitment to me I never would have had the opportunity or the strength to complete this project. Thank you for being my support system and for the constant encouragement.

ACKNOWLEDGMENTS

I would like to thank my dissertation co-chairs Dr. Marley Watkins and Dr. John Balles for providing me with guidance and constant mentorship throughout this process. The opportunity to learn from these two excellent scholars and practitioners has provided me with a wealth of knowledge that I look forward to bringing into my professional career in the future. In addition, I would like to thank Dr. Christa Lynch for participating on my committee and for her advisement throughout this project. I also owe a large thank you to Dr. Samuel Green for his continual statistical help and guidance. Likewise, a thank you to the multiple graduate students that spent hours in the special education record rooms collecting data that was eventually used in this project. Thank you to the many other family, friends and colleagues that provided guidance and assistance with the completion of this process. Finally, thank you to God for once again showing me that in Him all things are possible, even this!

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Chapter 1

INTRODUCTION

Of all psychological tests, standardized intelligence tests are some of the most widely used by psychologists (Wilson & Reschly, 1996). School psychologists in particular commonly use standardized intelligence tests as one component of a psychoeducational evaluation for the determination of special education eligibility (Suzuki & Valencia, 1997), especially for the diagnosis of specific learning disabilities (SLD) that affect approximately 5% of the school-aged population and comprise over 50% of the special education population (Anyon, 2009).

Historically, diagnosis and eligibility of SLD has depended on a discrepancy model in which a child's ability, as measured by standardized intelligence tests, is compared to his/her skill in a specific academic domain as measured by standardized achievement tests. Until the most recent reauthorization of the Individuals with Disabilities Education Act (2004), the use of a standardized intelligence test was required for the assessment of a SLD. Upon the reauthorization of IDEA other methods of identification were approved, such as evidence of a failure to respond to evidence based interventions. However, in practice, the discrepancy approach is still commonly utilized as a diagnostic approach for identifying learning disabilities in students (Kavale & Spaulding, 2008).

Unfortunately, there are many negative outcomes associated with special education. Children with disabilities have been found to fall behind their peers without disabilities on multiple measures of societal attainment (Phelps & Hanley-Maxwell, 1997). Those with disabilities are more likely to be delinquent, unemployed, and have a lower socio-economic status (Blackorby & Wagner, 1996). Special education services have been shown to have either a negative or a statistically non-significant effect on children's reading and math skills (Morgan, Frisco, Farkas, & Hibel, 2010). Additionally, special education services did not improve children's externalizing or internalizing problem behaviors (Morgan et al., 2010).

The use of the discrepancy approach when identifying learning disabilities in students necessitates the use of standardized intelligence tests; additionally, the diagnosis of mental retardation also requires the use of standardized intelligence tests. In accordance with the American Psychological Association's code of ethics (2002), the National Association of School Psychologists' code of ethics (2010), and the Joint Committee on Testing Practices (2004), it is expected that all psychologists use tests that produce interpretable scores that are reliable and valid. Reliability is defined as the degree to which test scores are consistent and stable across conditions (Reynolds, Livingston, & Wilson, 2009). The Standards for Educational and Psychological Testing (1999) define validity as "the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests" (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 1999, p. 9). As specific tests are developed and revised the validity of the individual test needs to be established or reestablished by “accumulating evidence to provide a sound scientific basis for the proposed score interpretations” (AERA, APA, & NCME, 1999, p.9).

The traditional view of validity discussed validity in terms of content, criterion, and construct validity (Cronbach & Meehl, 1955). Content validity was defined as items that sample all aspects of a universal principle; criterion validity was described as test items that are related to an external criterion that can be measured in a concurrent manner (such as a behavior rating) or a predictive manner (such as SAT scores and college GPA); and finally construct validity was specified as test items that correlate with the theoretical structure of the construct being measured. For many years this “holy trinity” was regarded as the best way to evaluate the validity of a test; however, Messick (1995) developed the Unitarian view of validity, which regards validity not as a property of the test (as the classical theory does) but as an argument, or an evaluative judgment, which one makes about the meaning of the test scores.

Messick (1995) described six aspects of construct validity that must be addressed to appropriately evaluate a measure: content, substantive, structural, generative, external, and consequential. The content aspect includes representativeness (inclusion of all aspects of the measured domain), content relevance, and sampling of tasks that represent all important parts of the construct (functional importance). The substantive aspect includes measurement of the theoretical foundation of the construct through both process modeling and empirical evidence, indicating that the theoretical processes expected to be a component of the construct are in fact evident and measured appropriately. The structural aspect indicates that the structure of the known construct should be consistent with what a test is measuring. The generalizability aspect addresses how well scores and score interpretations generalize across groups, raters, settings, time, and tasks. The external aspect represents the relationship between the test scores and other criteria that theoretically measure similar constructs (convergent) as well as the relationship between the test scores and other criteria that theoretically measure opposing constructs (discriminant); specifically, the test should have strong correlations with similar constructs and weak correlations with discriminant constructs. The final aspect, consequential, refers to evidence that the interpretations of scores are appropriate and not representative of any bias or unfairness. Using all aspects of validity one can develop an argument for validity and thus develop evidence to support a specific test. In accordance with the aforementioned guidelines, as well as best practice (NASP, 2008), it is essential that cognitive tests used during psychoeducational assessments, particularly those used in the decision making process for special education, be psychometrically sound (reliable and valid).

Empirical studies of the identified components of construct validity have been frequently conducted with the Wechsler scales of intelligence. Wechsler's original scale of intelligence, the Wechsler – Bellevue Intelligence Scale (Wechsler – Bellevue; Wechsler, 1939), was created for use with the adult population but within a decade was modified to allow assessment of children via the Wechsler Intelligence Scale for Children (WISC; Wechsler, 1949). The WISC was twice modified over the ensuing decades, first with the Wechsler

Intelligence Scale for Children-Revised (WISC-R; Wechsler, 1974) and next with the Wechsler Intelligence Scale for Children-Third Edition (WISC-III; Wechsler, 1991). All three WISC versions have been thoroughly researched (Sattler, 2008). Additionally, more recent research regarding evidence of construct validity has occurred with the Wechsler Intelligence Scale for Children-Fourth Edition (WISC-IV; Wechsler, 2003a). Given the multiple test revisions that have occurred with the WISC, it is essential to address evidence regarding the structural validity of the test; specifically, that the internal structure of the test corresponds with the structure of the theoretical construct being measured (Messick, 1995).

Intelligence as a Measurable Construct

Intelligence is considered to be a trait, indicating that it should be relatively stable across time (Hunt, 2011). Since cognitive tests are measuring a stable trait it is expected that good test-retest reliability should be evident (Wright, 2011). Research has indicated that cognitive test scores have remained fairly stable from about the age of 5 through adulthood (Chen & Siegler, 2000). Individual differences in general intelligence (*g)* have been shown to remain highly stable over time in both average and highly select samples (Reeve & Bonaccio, 2011; Simonton, 2011). For example, one longitudinal study of ability tests reported a test-retest stability coefficient of .66 across a 66 year interval with an estimated short term test-retest reliability of .90 (Deary, Whalley, Lemmon, Crawford, & Starr, 2000). These findings are all indicative of intelligence being a stable trait that should be replicable in a test for an individual over time. If intelligence is considered to be a stable trait, then any change in the factor structure of a test can be interpreted as a problem with the specific test and not as an underlying change in the measured construct. Therefore, if the structure of the test changes over time then the construct validity of the test is limited because test-retest score differences cannot be explicitly interpreted as reflecting changes in the underlying construct.

Previous Wechsler Intelligence Tests

The Wechsler series of intelligence tests has long been regarded as one of the most popular cognitive assessments among clinicians (Alfonso, Oakland, LaRocca, & Spanakos, 2000; Belter & Piotrowski, 2001; Pfeiffer, Reddy, Kletzel, Schmelzer, & Boyer, 2000). Wechsler initially defined intelligence as “the aggregate or global capacity of the individual to act purposefully, to think rationally, and to deal effectively with his environment” (Wechsler, 1939, p. 3). The Wechsler intelligence tests were not initially based on a specific theoretical perspective. Instead, Wechsler focused on creating a test that had content evenly divided between verbally loaded tasks and tasks that were primarily nonverbal (Zachary, 1990).

Wechsler chose to focus on verbal and nonverbal, or performance, tasks not because the subtests measured different types of intelligence but to measure intelligence in different ways (Wechsler, 1958). Wechsler recognized that many individuals were undoubtedly intelligent but appeared to have low verbal abilities and thus it was essential to measure both ways of expressing intelligence. Wechsler’s overall purpose was to measure performance as a whole or general intelligence, not to measure specific abilities in isolation (Zachary, 1990). Continued revision of the Wechsler scales has increased the domains of cognitive functioning measured. Specifically, the measurement of more discrete domains of intelligence, such as processing speed and working memory, have been added to better identify a person's overall cognitive ability. As the Wechsler scales have continued to be revised, concern regarding the structural validity of each version of the test has been apparent and has been a focus of study. Structural validity has been examined in the Wechsler scales by using exploratory factor analysis (EFA). EFA is a statistical method that is used to help develop theories and better understand how theoretical constructs are structured.

The Wechsler Intelligence Scale for Children (WISC; Wechsler, 1949) was composed of 12 subtests similar to the Wechsler-Bellevue Intelligence Scale, but modified to be age appropriate. These 12 subtests were used to generate Verbal IQ (VIQ), Performance IQ (PIQ), and Full Scale IQ (FSIQ) scores. Research regarding the structural validity of the WISC is scarce; however, Cohen (1959) examined the factor structure of the WISC at three age groups (7.5, 10.5, and 13.5 years). Using Thurstone's complete centroid method (1947), a five-factor structure was identified with a second-order general factor, *g*. The five factors identified were Verbal Comprehension I (Information, Similarities, and Vocabulary), Perceptual Organization (Block Design, Object Assembly, and Picture Completion), Freedom from Distractibility (Digit Span and Arithmetic), Verbal Comprehension II (Comprehension, Vocabulary, and Picture Completion), and Quasi-Specific (Coding and Picture Arrangement). When this same data was reanalyzed using different techniques, such as examining the average proportion of the total variance attributable to unrotated factors and the Kaiser (1960) method of retaining as many factors as latent roots greater than one, only two identifiable factors (Verbal Comprehension and Perceptual Organization) were identified (Silverstein, 1969).

The Wechsler Intelligence Scale for Children – Revised (WISC-R; Wechsler, 1974) was composed of the same subtests and IQ scores as the WISC; however, adjustments were made to the age range (6 to 16 years rather than 5 to 15 years) appropriate for this measure. The structure of the WISC-R consisted of the same two factors: Verbal Comprehension (VC) and Perceptual Organization (PO). Additionally, the subtests included in each factor were the same as in the WISC. Using the normative sample, Wallbrown, Blaha, Wallbrown, and Engin (1975) found strong support of a hierarchical factor structure of the WISC-R. Kaufman (1975) also examined the normative sample using principal-factor analysis with varimax rotation of two-, three-, four-, and five-factor solutions across each age level. He found that for six age groups there were two identified factors (VC and PO), but at the remaining age groups (8.5, 10.5, 13.5, and 15.5 years) the three-factor structure (VC, PO, and Freedom from Distractibility or FD) was most appropriate. The additional FD factor included the Arithmetic, Digit Span, and Coding subtests. Overall, Kaufman (1975) argued that his results were supportive of a two-factor solution (VC and PO) as identified by Wechsler's divisions of subtests. In contrast, McMahon and Kunze (1981) found that the three-factor solution was appropriate for exceptional children. The third factor appeared to differ across samples and the interpretation of the factor remained unclear (Zachary, 1990). However, as additional clinical studies were conducted the two-factor solution remained the most stable. This factor solution has been shown to be relatively invariant across age (Conger, Conger, Farrell, & Ward, 1979); sex (Reynolds & Gutkin, 1980); ethnicity (Dean, 1980; Gutkin & Reynolds, 1980, 1981; Reschly, 1978); and psychiatric diagnoses (Petersen & Hart, 1979).

Revision of the WISC-R produced the Wechsler Intelligence Scale for Children-Third Edition (WISC-III; Wechsler, 1991), which included the same subtests as the WISC-R with the addition of one subtest, Symbol Search. However, the VIQ and PIQ scores were dropped in favor of Verbal Comprehension Index (VCI), Perceptual Organization Index (POI), Freedom from Distractibility Index (FDI), and Processing Speed Index (PSI) scores. The reported factor structure of the normative sample included a second-order general ability factor, *g*, (Spearman, 1904) and four first-order factors corresponding to the index scores: Verbal Comprehension (Information, Similarities, Vocabulary, and Comprehension), Perceptual Organization (Picture Completion, Picture Arrangement, Block Design, and Object Assembly), Freedom From Distractibility (Arithmetic and Digit Span), and Processing Speed (Coding and Symbol Search). The WISC-III factor structure was subsequently investigated in independent samples. Roid, Prifitera, and Weiss (1993) analyzed the factor structure of the WISC-III with a nationally representative sample (*n* = 1,118). Through the use of multiple criteria in identifying the number of factors, they replicated the four-factor structure found with the normative sample. Additional analysis conducted with the Canadian normative sample (*n* = 1,100) also confirmed the four-factor structure (Roid & Worrall, 1997). The four-factor structure was also found to be the best solution in clinical samples of psychiatric inpatients (Tupa, Wright, & Fristad, 1997) and children identified as eligible for special education services (Konold, Kush, & Canivez, 1997; Grice, Krohn, & Logerquist, 1999). There have been multiple critiques of the four-factor model for the WISC-III (Carroll, 1993; Sattler, 1992). These critiques were typically due to the smaller third and fourth factors (FD and PS). However, in general the four-factor structure of the WISC – III normative sample has been accepted (Grice et al., 1999).

Wechsler Intelligence Scale for Children-Fourth Edition

The Wechsler Intelligence Scale for Children-Fourth Edition (WISC-IV; Wechsler, 2003a) was significantly different from the WISC-III. For example, it was developed in alignment with current intelligence theories (Wechsler, 2003b), specifically the Cattell-Horn-Carroll (CHC; McGrew & Flanagan, 1998) theory of intelligence. The CHC theory of intelligence regards intellectual abilities within a hierarchical structure consisting of three strata (general ability, broad abilities, and narrow abilities). Specifically, general intellectual ability, *g*, is represented on stratum III and ten broad cognitive abilities are represented on stratum II: Fluid Intelligence (*Gf*), Crystallized Intelligence (*Gc*), Quantitative Knowledge (*Gq*), Reading and Writing (*Grw*), Visual Processing (*Gv*), Auditory Processing (*Ga*), Short-term Memory (*Gsm*), Long-term Storage and Retrieval (*Glr*), Processing

Speed (*Gs*), and Decision/Reaction Time/Speed (*Gt*). Finally, 70 narrow abilities are included on stratum I.

According to Wechsler (2003b), the alignment of the WISC-IV to the CHC theory of intelligence resulted in the creation of new subtests as well as the removal of existing subtests. Changes to the subtest structure of the WISC-IV included the addition of five subtests (Word Reasoning, Picture Concepts, Matrix Reasoning, Letter-Number Sequencing, and Cancellation), making the Information subtest supplemental, and the removal of three WISC-III subtests (Picture Arrangement, Object Assembly, and Mazes). Additionally, revisions occurred at the item level across subtests and approximately 60% of items in the core subtests were new or revised (Watkins, 2010). The WISC-IV contains 15 subtests (10 core and 5 supplementary). The 10 core subtests include Block Design, Similarities, Digit Span, Picture Concepts, Coding, Vocabulary, Letter-Number Sequencing, Matrix Reasoning, Comprehension, and Symbol Search. The 5 supplementary subtests include: Picture Completion, Cancellation, Information, Arithmetic, and Word Reasoning. The index scores are identified as the Verbal Comprehension Index (VCI), Perceptual Reasoning Index (PRI), Working Memory Index (WMI), and Processing Speed Index (PSI).

Psychometric properties of the WISC-IV were determined using the normative sample of 2,200 children (Wechsler, 2003b). The overall sample yielded average internal consistency reliability coefficients across subtests ranging from .79 (Symbol Search and Cancellation) to .90 (Letter Number Sequencing), with all other subtests between .80 and .89. Importantly, the reliability coefficients of the WISC-IV subtests are substantially larger than the corresponding subtests on the WISC-III, indicating that the WISC-IV has better reliability than the previous edition. The internal consistency reliability coefficients for composite scores were .88 for Processing Speed, .92 for Perceptual Reasoning and Working Memory, .94 for Verbal Comprehension, and .97 for Full Scale IQ.

A special group was formed consisting of 661 exceptional children split into the following groups: Intellectually Gifted; Mental Retardation – Mild Severity; Mental Retardation – Moderate Severity; Reading Disorder; Reading and Written Expression Disorders; Mathematics Disorder; Reading, Written Expression, and Mathematics Disorders; Learning Disorder and Attention-Deficit/Hyperactivity Disorder; Attention – Deficit/Hyperactivity Disorder; Expressive Language Disorder; Mixed Receptive-Expressive Language Disorder; Open Head Injury; Closed Head Injury; Autistic Disorder; Asperger's Disorder; and Motor Impairment. Approximately 5.7% of the normative sample was composed of children from this special group. Reliability coefficients were calculated for the special groups in the same manner as with the standardization sample. For subtests, the special groups sample yielded average internal consistency reliability coefficients ranging from .82 (Digit Span Forward) to .93 (Letter – Number Sequencing and Matrix Reasoning). These results indicate that the WISC-IV is an equally reliable measure for the cognitive assessment of children from the general population as well as exceptional children (those with clinical diagnoses). The overall internal consistency reliability coefficients for special groups across composite scores were not included in the technical manual (Wechsler, 2003b).

The normative sample was also utilized to confirm evidence of test-retest stability. Using 243 children, participants were twice administered the WISC-IV with an interval between test and retest ranging from 13 to 63 days. Overall results indicated that the scores remained stable across all age groups; however, there appeared to be practice effects due to the short interval of time between test administrations (Wechsler, 2003b). Research has found that when the test-retest interval exceeds one year, practice effects are not typically observed or are so small that it does not significantly affect the stability coefficients (Ryan, Glass, & Bartels, 2010).

Structural Validity of the WISC-IV

Normative sample. Exploratory factor analysis (EFA) was used to examine the factor structure of the WISC-IV with the normative sample (Wechsler, 2003b). Using the 10 core subtests of the WISC-IV, an EFA analysis found that the four-factor theoretical model was appropriate, with each subtest loading primarily on its predicted factor. The EFA validated the following factor structure: Verbal Comprehension (Similarities, Vocabulary, and Comprehension), Perceptual Reasoning (Block Design, Picture Concepts, and Matrix Reasoning), Working Memory (Digit Span and Letter – Number Reasoning), and Processing Speed (Coding and Symbol Search). These initial findings indicated that in the youngest age group (ages 6 – 7 years) Picture Concepts loaded evenly on the Perceptual Reasoning and Verbal Comprehension factors; however, this was only evident within one age group and Picture Concepts loaded primarily on Perceptual Reasoning in the other age groupings.

Confirmatory factor analysis (CFA) studies were also applied to the normative sample to better understand the structure of the WISC-IV (Wechsler, 2003b). CFA is a statistical method that allows for the investigation of relationships between measured variables and the underlying hypothetical constructs (Tabachnick & Fidell, 2007). CFA differs from EFA because it is typically used to test theory rather than to develop theories (Keith, 2005). CFA models have multiple components: factors, indicators, and measurement error. "Each indicator is a continuous variable represented as having two causes – a single underlying factor that the indicator is supposed to measure and all other unique sources of causation that are represented by the error term" (Kline, 2005, p. 166). The indicators are measured variables that have direct relationships with a factor; these direct effects are measured by statistical estimates, typically regression coefficients, and are called factor loadings (Kline, 2005).

CFA is used to further investigate proposed models as well as to test theories. The simplest form of CFA stipulates that the nature of the factor structure underlying the data is determined in advance. The researcher specifies the number of factors, which variables will load on each factor, and if factors are correlated or uncorrelated (Keith, 2005). The goal of a CFA is to "confirm" that the hypothesized model is a good explanation of the data. Once a model is proposed, CFA applies the model to the data sample. The results of the sample data are then compared to the hypothesized structure that is expected to be found in the population and the difference between the models is assessed. This analysis results in fit statistics, which indicate how well the overall model (identified factor structure) fits the sample data. Importantly, values of fit are not indicating that results have theoretical meaning, simply that the proposed model fits the data (Kline, 2005). Using the fit statistics, the researcher is able to examine the accuracy of a specific factor structure by applying constraints to the solution and determining if the more restricted solution remains consistent with the data sample.

There are many different fit statistics, or indices, described in the literature (Kline, 2005). The most common goodness of fit measure is the chi-square statistic (χ^2); which is used in conjunction with the degrees of freedom (*df*), which measure the degree to which a model is over-identified. A small χ^2 combined with a large *df* indicates statistical insignificance ($p > .05$), and thus the model fits the data. This index is actually considered to be a "badness of fit" index because the model's fit is worse when the χ^2 value is high. The χ^2 fit statistic is sensitive to sample size. Specifically, large samples typically result in an underestimation of model fit and in small samples fit may be overestimated (Keith, 2005). Unfortunately, the χ^2 statistic assumes perfect population fit of the model and it is unlikely that any model will perfectly fit the data. Thus, although the most commonly reported index, it is not ideal.

Unlike the χ^2 index, the Tucker-Lewis Index (TLI; Tucker & Lewis, 1973) does not assume perfect population fit and appears to be robust in large and small samples. A value of $\geq .95$ on the TLI is demonstrative of good fit between the theoretical model and the sample data and a value of ≥ .90 is indicative of reasonable fit (Hu & Bentler, 1999). The root mean square error of approximation (RMSEA; Steiger, 1990) is a measure of approximation of fit rather than exact fit (as is the χ^2 index). RMSEA is also considered to be a "badness of fit" index as a higher number is not indicative of good fit. A value of ≤ .06 for the RMSEA is demonstrative of good fit between the theoretical model and the sample data (Hu & Bentler, 1999) and a value ≤ .08 suggests reasonable fit (Browne & Cudeck, 1993). These criteria are routinely reported, yet, not all researchers agree that these criteria provide enough information for decision making. Bollen and Long (1983) indicated that, "The test statistics and fit indices are very beneficial, but they are no replacement for sound judgment and substantive expertise" (p. 8).

Wechsler (2003b) investigated multiple hypothesized structural models for the WISC-IV ranging from one to four-factor models. Goodness of fit indices indicated that the four-factor model was the best fit as compared to the null model across all age groups and for the overall sample (χ^2 = 131.62 (29); TLI = .98, RMSEA = .04). When conducting exploratory and confirmatory factor analyses with the normative sample, Wechsler (2003b) failed to evaluate a multi-level structure for the WISC-IV. Hierarchical CFA models are used when there are hierarchical relationships within the underlying theoretical constructs. An indirect hierarchical model is when a second-order factor has a direct effect on the first-order factors and an indirect effect (through the first-order factors) on the indicators (Kline, 2005). With intelligence testing, *g* is regarded as a higher-order factor because it is indirectly measured by other factors (such as Verbal and Perceptual). An additional model that has been explored within intelligence testing is the direct hierarchical model. A direct hierarchical model allows the general intelligence factor to have a direct effect on the individual subtests and each of the first order factors to have a direct effect on its specific subtests with no indirect effects (Gignac, 2008). There have been subsequent studies completed by independent investigators to correct this omission.

The first to explore hierarchical models for the WISC-IV was Keith (2005), who utilized the normative sample and CFA methodologies to investigate the hierarchical factor structure of all 15 subtests of the WISC-IV. Two of the models that were tested included a general intelligence factor; both of these models indicated good fit to the data. Initially, an indirect hierarchical model (Gignac, 2008) was specified. In an indirect hierarchical model, it is assumed that the subtests are best explained by the first-order factors (VC, PR, WM, and PS) and the first-order factors are best explained by the second-order factor, general intelligence (see Figure 1 for an illustration).

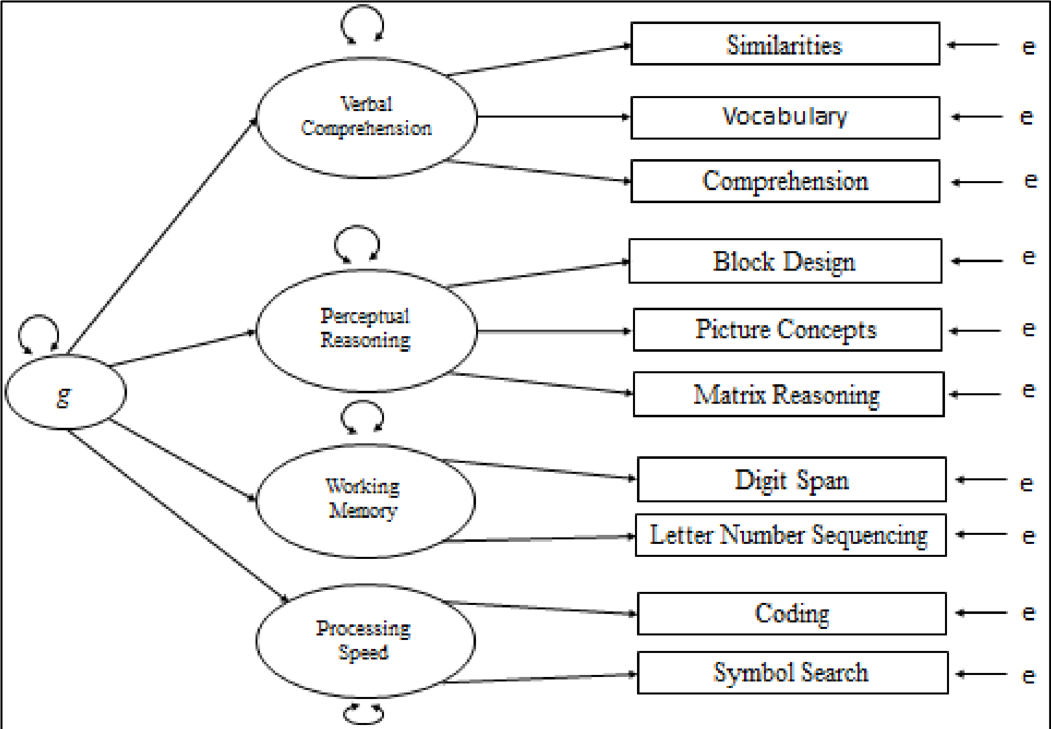


Figure 1. An indirect hierarchical structure of the WISC-IV.

This model displayed good fit to the normative sample data (RMSEA =.05, χ^2 = 296.93 (86), CFI = .97, SRMR = .04). Keith (2005) also specified a direct hierarchical model as an alternative to the indirect hierarchical model. A direct hierarchical model allows all subtests to directly load on the first-order factors (VC, PR, WM, and PS) as well as on the second-order factor (*g*). Using this model there is no assumed relationship between first- and second-order factors (see Figure 2 for an example). This model exhibited better fit than the indirect hierarchical model (RMSEA = .04, χ^2 = 202.6 (75), CFI = .98, SRMR = .03).

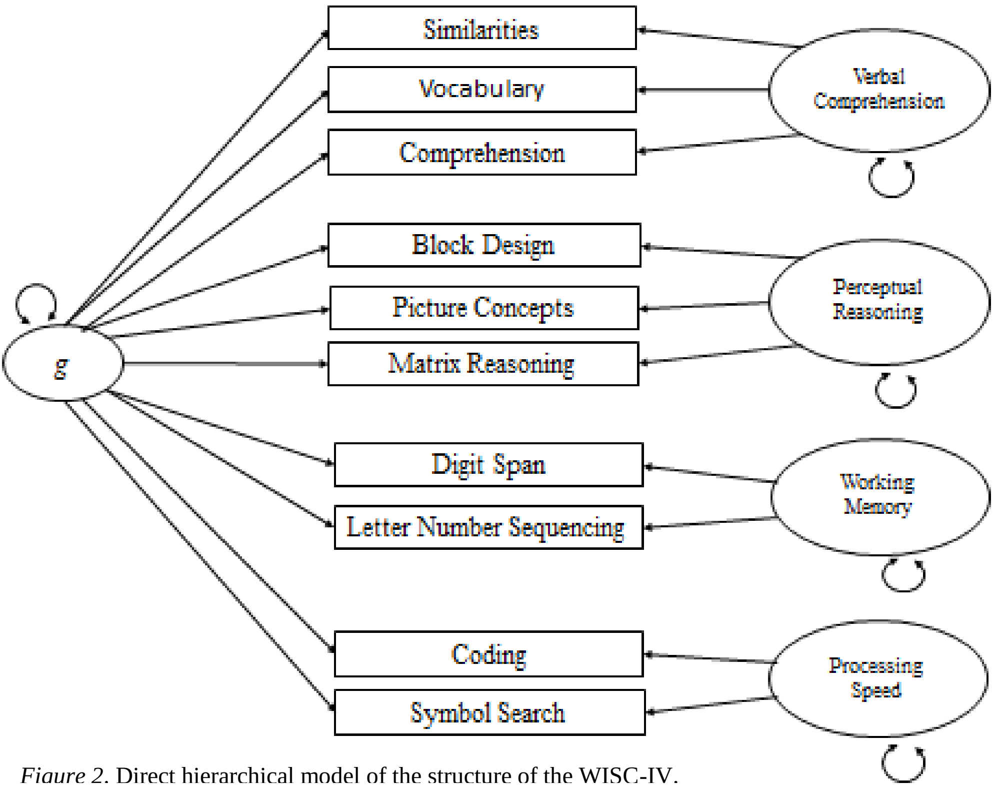


Figure 2. Direct hierarchical model of the structure of the WISC-IV.

Keith, Fine, Taub, Reynolds, and Kranzler (2006) also used the normative sample to investigate the factor structure of the WISC-IV. They found that the hypothesized factor structure of the WISC-IV, according to the technical manual, was not the best fitting model. Specifically, they found that by imposing an indirect hierarchical factor model that theoretically underlies the WISC-IV (*g* as the second order factor), the model fit worsened. This indicated to them that the factor model proposed by Wechsler (2003b) is not a good explanation of the constructs measured. As an alternative, they hypothesized that a theoretically derived structure based on the Cattell-Horn-Carroll Model (CHC; McGrew, 1997) would better describe the abilities measured by the WISC-IV. The model based on CHC theory did yield a better fitting model with the standardization data than the four-factor theoretical model identified in the WISC-IV technical manual. However, this higher order five-factor structure was the best fitting model only when utilizing all 15 of the subtests available on the WISC-IV (including the core and supplemental subtests). Thus, it may not hold when only the core subtests are considered. Clinicians traditionally exclusively use the core subtests and are unlikely to administer the supplemental subtests (Watkins, 2010). By using all 15 subtests the clinical utility of the higher order five-factor structure is substantially reduced.

There were additional limitations to the CHC model specified by Keith et al. (2006) for the WISC-IV normative sample. For example, this model abandoned simple structure and allowed cross loadings. That is, subtests were allowed to load on more than one factor. Permitting subtests to cross load creates difficulty in understanding the resulting factor scores. As the subtests were created to measure specific areas of ability, this limits the clinical utility of the information provided. Additionally, this analysis reported that the loading of the second-order factor (*Gf*) on the third-order general factor (*g*) was 1.00, indicating dependence of these two factors. Allowing these two factors to be dependent indicates that the *Gf* factor was not necessary for model fit. This is problematic because the *Gf* factor is an essential component of the CHC theory of intelligence.

Additional analysis using the nationally representative standardization sample was conducted by Watkins (2006), who used the Schmid and Leiman (1957) orthogonalization procedure to evaluate the factor structure of the WISC-IV. It was determined that the WISC-IV general factor accounted for approximately 38.3% of the total variance in the core subtests. The verbal comprehension factor (VC) explained an additional 6.5% of the total variance; the perceptual reasoning factor (PR) explained an additional 2.2% of the total variance; the working memory factor (WM) explained an additional 2.3% of the total variance; and the processing speed factor (PS) explained an additional 4.4% of the total variance. These findings were similar to previous research conducted on earlier revisions of the Wechsler intelligence scales (Gustafsson & Undheim, 1996). Overall, Watkins (2006) found that the general factor explained more variance than any of the first order factors and suggested that the FSIQ is the best predictor of intellectual ability.

Clinical samples. Utilizing the stratified normative sample for structural analyses has limitations. The largest limitation is the exclusion of specific subsets of the population; specifically, a lack of structural validity evidence for disabled/exceptional/clinical populations. The Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999) specify that validity of constructs must be established in the population for which measures are created. As the WISC-IV is most commonly applied with an exceptional population (those with clinical diagnoses), it is imperative that the factor structure of the WISC-IV be validated across clinical populations in addition to the normative sample.

Accordingly, several studies have been conducted with clinical samples. The first study included a sample of 432 students referred for evaluation for special education eligibility (Watkins, Wilson, Kotz, Carbone, & Babula, 2006). Of these participants, 65% were identified as eligible for special education under the following categories: learning disabilities (37%), gifted (8%), emotional disabilities (7%), mental retardation (5%), multiple disabilities (6%), and speech disabilities (2%). The researchers used a four-factor EFA with Schmid-Leiman (1957) orthogonalization, which permits the variance accounted for by the higher-order factor to initially be extracted, followed by the residual variance accounted for by the group factors. Results indicated that a four-factor solution had excellent fit, accounting for 62% of the total variance. Additionally, the general intelligence factor explained 47% of the total variance whereas the first order factors accounted for significantly less variance (ranging from 1.4% (working memory) to 6.5% (verbal comprehension) of total variance. Overall, this study found that the proposed four-factor model (Wechsler, 2003b) was appropriate for the referred sample of this study. Moreover, it determined that the general intelligence factor accounted for a greater amount of total variance than the first-order factors and thus the authors did not recommend interpretation of the first-order factor scores over the reported general intelligence score.

More recently, Bodin, Pardini, Burns, and Stevens (2009) conducted a CFA to examine the higher order factor structure of the WISC-IV in a clinically referred sample (*N* = 344, 217 males, *M* age = 10.4 years). The sample consisted of children with the following diagnoses: attention deficit/hyperactivity disorder (20%), epilepsy (18%), learning disability (14%), traumatic brain injury (9%), cerebral palsy (4%), meningitis/encephalitis (3%), spina bifida (2%, in-utero/perinatal conditions (1%), and other medical conditions (29%). CFA's were used to replicate the models tested in the normative sample (one-factor, two-factor, three-factor, and four-factor), and each of the multiple factor models included a second order factor representing general intelligence (*g*). Results indicated that the indirect hierarchical four-factor model was preferred. Overall, the general intelligence factor explained the most variance (48.3% of total variance); whereas, the first order factors accounted for significantly less variance: verbal comprehension (5.2% of total variance), perceptual reasoning (2.5% of total variance), working memory (0.2% of total variance), and processing speed (6.4% of total variance). These findings are consistent with previous research regarding both the normative sample as well as referred samples (Wechsler, 2003b; Watkins, 2006; Watkins, et al., 2006).

Most recently, Watkins, (2010) investigated the structure of the WISC-IV in a national sample of children referred for psychoeducational evaluations (*N* = 355 students, 218 males; *M* age = 9.78 years). The sample consisted of children with the following diagnoses: learning disability (41%), other health impairments (9%), mental retardation (7%), emotional disabilities (6%), speech disabilities (4%), gifted (2%), and autism spectrum disorders (1%). Around 30% of the participants were not found to have a disability. CFA methods were used with maximum likelihood estimation in order to evaluate six hypothesized structural models of the WISC-IV (one-factor, two-factor, three-factor, four-factor, indirect hierarchical, and direct hierarchical) (See Figures 3 – 6 for examples of one-, two-, three- and four-factor models).

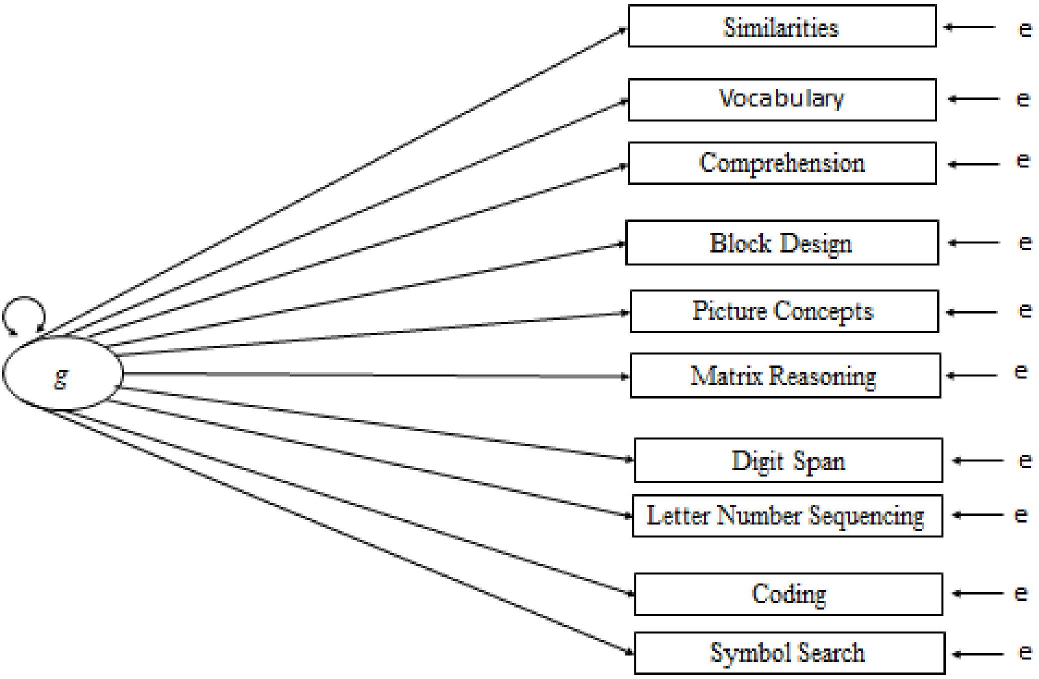


Figure 3. A one-factor model of the WISC-IV structure.

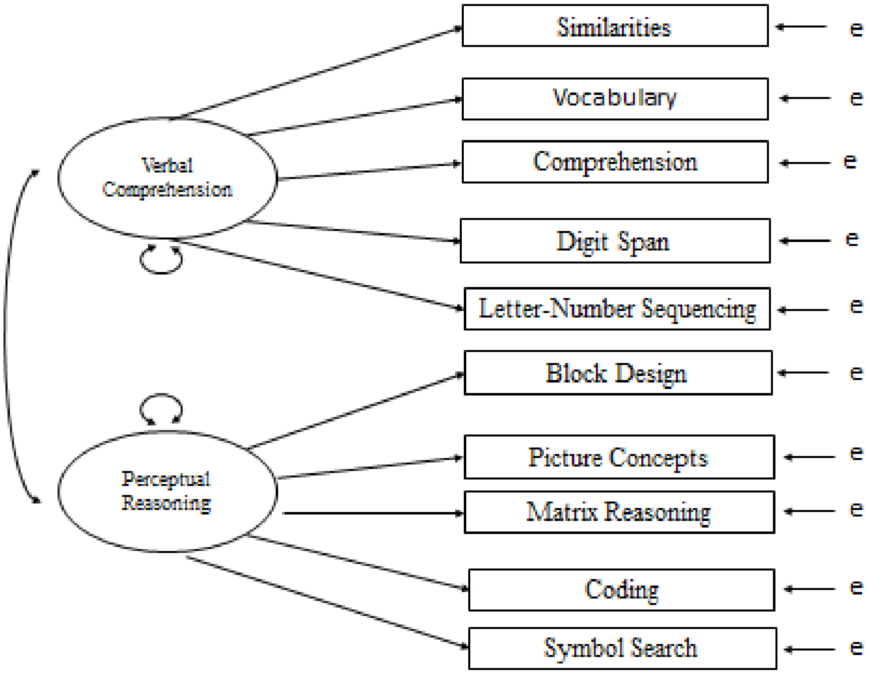


Figure 4. A two-factor oblique model of the structure of the WISC-IV.

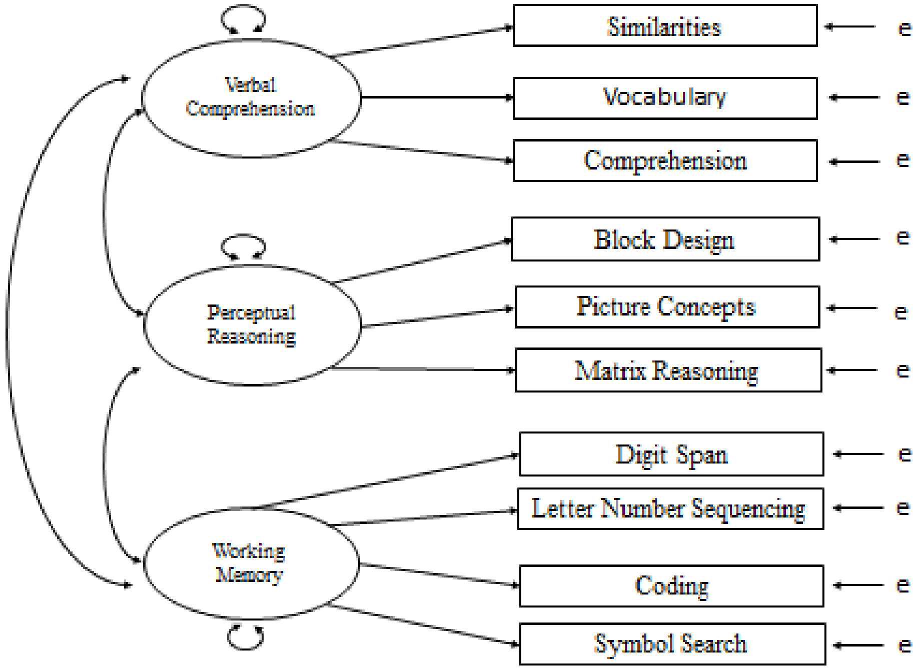


Figure 5. A three-factor oblique model of the structure of the WISC-IV.

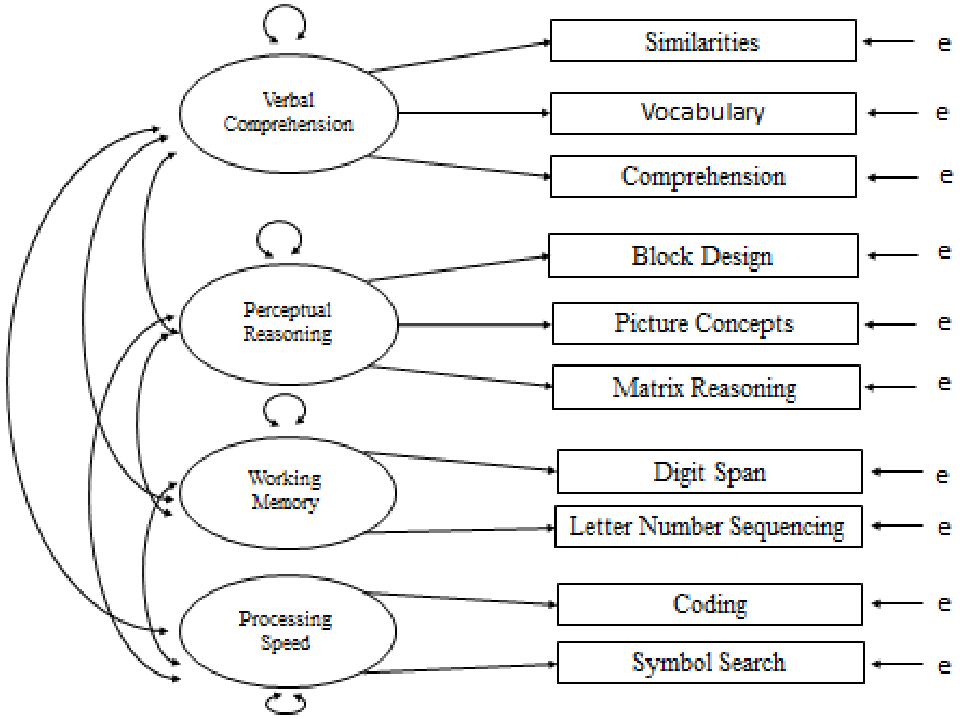


Figure 6. A four-factor oblique model of the structure of the WISC-IV.

The one-, two-, and three-factor models did not exhibit good fit, but the other three models did: the four-factor oblique model (RMSEA = .06, SRMR = .028, CFI = .981), indirect hierarchical model (RMSEA = .058, SRMR = .03, CFI = .98), and the direct hierarchical model (RMSEA = .058, SRMR = .028, CFI = .983). Watkins (2010) concluded that "the WISC-IV general intelligence factor is best interpreted as a first-order breadth factor as specified in the direct hierarchical model" (p. 786) and determined that the first-order four-factor model favored by Wechsler (2003b) was not appropriate because it did not include general intelligence as required by the theoretical structure of the WISC-IV. The direct hierarchical model was superior statistically to the indirect hierarchical model (df = 2, $\Delta\chi^2$ = 6.68, p =.048). For the direct hierarchical model, the general intelligence factor explained the most variance (47% of total variance); whereas, the first-order factors accounted for significantly less variance: verbal comprehension (4.8% of total variance), perceptual reasoning (3.1% of total variance), working memory (1.9% of total variance), and processing speed (6.1% of total variance). Overall, it was found that the general intelligence factor was best interpreted by the direct hierarchical model (see Table 1 for a comparison of variance components found across factor analyses of the WISC-IV).

Table 1.

Comparison of Total Variance Components for First- and Second- Order Factors Across Studies

Study	General Intelligence (*g*)	Verbal Comprehension (VC)	Perceptual Reasoning (PR)	Working Memory (WM)	Processing Speed (PS)
Watkins (2006)	38.3%	6.5%	2.2%	2.3%	4.4%
Watkins, et al. (2006)	46.7%	6.5%	2.4%	1.4%	4.7%
Bodin, et al. (2009)	48.3%	5.2%	2.5%	0.2%	6.4%
Watkins (2010)	47%	4.8%	3.1%	1.9%	6.1%

Summary of WISC-IV structural validity evidence. EFA and CFA were used to examine the factor structure of the WISC-IV with the normative sample (Wechsler, 2003b). Both sets of analyses validated the four-factor structure (VC, PR, WM, and PS). However, Wechsler (2003b) did not investigate the multi-level structure of the WISC-IV. This oversight was corrected by subsequent researchers. Direct and indirect hierarchical models were examined and both indicated good fit, with the direct hierarchical model exhibiting better fit (Keith, 2005). Subsequent research by Keith, et al. (2006) indicated that using a five-factor model, based on CHC theory, yielded a better fitting model than the four-factor model identified by Wechsler (2003b). However, there were multiple limitations to this study. Additional analysis of the normative sample determined that the general factor explained the most variance overall (Watkins, 2006).

Multiple studies also investigated the structure of the WISC-IV in clinical samples. A four-factor EFA with Schmid-Leiman (1957) orthogonalization verified that the four-factor model proposed by Wechsler (2003b) was appropriate for a sample of students referred for evaluation for special education eligibility (Watkins, et al., 2006). Subsequent research with a clinical population indicated that the higher order four-factor model was the most appropriate model (Bodin et al., 2009). Finally, a study consisting of children referred for psychoeducational evaluations examined multiple structural models of the WISC-IV and determined that four-factor, indirect hierarchical, and direct hierarchical models all displayed good fit, but the direct hierarchical model best explained the general intelligence factor (Watkins, 2010).

Structural Validity of WISC Across Time

Previous research studies have evaluated the structure of the WISC-IV by conducting cross-sectional studies using subjects with ages from 6 to 16 years. That is, the structure found with children of a specific age was compared to different children of other ages. Cross-sectional studies allow the researcher to investigate many participants of different ages at one time. The researcher is then able to make comparisons between ages. However, a major limitation of this design is that there are cohort effects. A cohort effect is the variation that occurs between groups based on differences due to possible shared temporal experiences (such as year of birth, year the child began school, historical significance, etc.). The ability to study the same sample of participants across time, referred to as a longitudinal design (Rosenthal & Rosnow, 1991), is a more time – consuming, yet desirable, method to investigate the change of individuals over time because it controls for any variation between groups by sampling the same individuals multiple times.

Studies of cognitive assessment measures have typically used cross-sectional designs to determine that the structure of the test remained constant across age groups. For example, the study by Keith et al. (2006) of the WISC-IV examined the normative sample to determine if the WISC-IV subtests measure the same constructs across age groups. Using multisample CFA models, Keith et al. (2006) constrained the variance and covariances to be equal across age groups and determined that this model had good fit (RMSEA = 0.05, TLI = 0.967). There was little difference, on average, between the actual correlations of the WISC-IV subtests and the predicted correlations from the hypothesized model. These findings indicated that the WISC-IV measures the same constructs across age groups. Thus, Keith et al. demonstrated that the factor structure of the WISC-IV was similar for a large group of children aged 6 through 16 years of age but did not demonstrate that the factor structure of the WISC-IV was similar for the same group of children as they matured across time.

Longitudinal factor analyses of the WISC. There have only been four longitudinal factor analyses of WISC scores across the past 45 years. In the first, the WISC factor structure was investigated with a sample of 153 pre-school aged children who were administered the WISC and followed up one year later with an additional administration of the WISC (Osborne, 1965). Using an EFA with varimax rotation, the factor structure changed from pre-school to first grade. Specifically, there were 8 factors for the time 1 administration and 10 factors for the time 2 administration. However, this study included children that were not of appropriate age for the WISC. Additionally, the methodology of this study is problematic as the subtests were split into two, three, or four parts to create additional variables and the EFA methods were sub-optimal (Gorsuch, 2003). Because of these limitations, the results of this study should be regarded with caution. Similar techniques and results were reported by Osborne, Anderson and Bashaw (1967) for the WISC with the same fatal limitations.

In the third study the WISC-R factor structure was examined using a longitudinal design with a sample (*N* = 322) of children eligible for special education services across a span of approximately 3 years (Juliano, Haddad, & Carroll, 1988). This study enrolled children who were identified as either white or black; other ethnicities were not included. Results indicated that for students who were administered the Digit Span subtest at Time 1 and Time 2 (*n* = 229), a three-factor solution was identified for all groups. The three known factors were: Verbal, Perceptual, and Freedom from Distractibility. Coefficients of congruence were used to quantify similarity between groups, and indicated that the three-factor solution remained stable for children with learning disabilities across the three-year time span regardless of sex or ethnicity.

The fourth longitudinal factor analysis investigated the factor structure of the WISC-III with 177 students classified as a child with a specific learning disability (SLD), a serious emotional disability (SED), mental retardation (MR), or other disabilities (Watkins & Canivez, 2001). These students were twice administered the WISC-III approximately 3 years apart. Four models were initially evaluated using CFA and the first-order, four-factor model was accepted as the best fitting model for both test and retest occurrences. Test and retest data was also analyzed to test for invariance of the factor structure across time. Initially, all factor loadings, factor variances, factor covariances, and subtest error variances were constrained to be equal; however, this model had inferior fit in comparison to a baseline model (χ^2=170 (126), *p* = .06). This was likely due to the error variances for three subtests (Vocabulary, Coding, and Arithmetic). Upon releasing those constraints, the model fit was significantly improved (χ^2=148.5 (123), *p* = .058). These results indicated that the WISC-III measured the same constructs across time and that the constructs were manifested in the same way across groups.

Using CFA for Analysis of Longitudinal Factor Structures

Previous studies using EFA methods have provided important yet incomplete information regarding invariance of factor structures. When using EFA methods the researcher must decide how many factors to retain, what method of extraction to use, and what method of rotation to apply. Upon analysis, the researcher combines theory, previous research, and the current findings to assign names to factors based on the specific factor loadings (Keith, 2005). When conducting invariance studies, EFA may allow for different factor structures across groups (or time). For example, item specific variances may result in a factor in one group (i.e., Time 1) and not in the other group (i.e., Time 2). However, by using CFA methods the researcher specifies the exact model that best explains the factor structure of the data. In longitudinal invariance studies this is especially important as the models at Time 1 and Time 2 are constrained to have the same number of factors, equivalent factor loadings, and intercepts at test and retest (Wu, Li, & Zumbo, 2007). Although an important variant of factor analysis, EFA involves a large amount of judgment, whereas CFA allows for comparison between specific models (based on fit indices) as they become more constrained. Thus, CFA should be the method of choice for testing invariance of equivalent models (Brown, 1996).

CFA methods are commonly used to analyze the longitudinal factor structure of tests (Stein, Lee, & Jones, 2006). These analyses typically begin by investigating whether the same measured variables define each factor at both test and retest occasions. If they do not, the test is not measuring similar construct(s) at test and retest occasions and test scores cannot therefore be meaningful compared across time. This is generally considered to be the least restrictive test of similarity of factors across time and has been called configural invariance (Chen, 2007). If configural invariance is found it posits that the overall factor pattern is the same at test and retest.

Even if the same measured variables define each factor at both test and retest occasions, they may not do so with equivalent precision. If, on the other hand, each measured variable loads equally on its corresponding factor at test and retest occasions then the same constructs are being measured with equal precision at both occasions. This has been labeled weak factorial invariance (or metric invariance) and indicates that the factors have the same meaning across time (Byrne, 2006; Widaman & Reise, 1997). Logically, this analysis is conducted after determining that configural invariance holds. Failure to achieve metric invariance indicates that the factor structure cannot be assumed to remain stable across time and therefore interpretation of change in test scores cannot be unequivocally attributed to change in the constructs being measured. Nor can test scores at test and retest be compared to other variables because one unit of change in test scores would not be equal to one unit of change in retest scores (Chen, Sousa, & West, 2005).

Configural and weak factorial invariance still allow factor means to differ across test and retest occasions. Similar to use of the Kelvin temperature scale for the test occasion and the Celsius scale for the retest occasion, the two scales can be correlated but their means differ. Thus, the factor intercepts must be tested and found to be equivalent before factor means can be compared. This level of invariance has been called strong factorial invariance (Widaman & Reise, 1997). When strong factorial invariance "is achieved, it means that scores from different groups [or two tests from the same group across time] have the same unit of measurement (factor loading) as well as the same origin (intercept), and thus the factor means can be compared across groups [or across time]. Otherwise, it cannot be determined whether any difference between groups on factor means is a true group difference or a measurement artifact" (Chen et al., 2005, p. 475). Logically, this analysis is conducted after determining that configural and weak factorial invariance holds.

Configural, weak, and strong factorial invariance still allow the error variances of measured variables to differ across test and retest occasions. This level of invariance has been referred to as strict factorial invariance (Widaman & Reise, 1997). When this level of invariance is achieved, it indicates that all differences between test and retest scores are solely due to group differences associated with the common factors. If strict factorial invariance is not met then it cannot be assumed that unique error variances are not contributing to differences between groups. Although some researchers have indicated that testing for the equality of error variances is the least important aspect of factorial invariance and thus not essential (Bentler, 2005), others have suggested that it is important to consider (Wu, Li, & Zumbo, 2007). Wu et al. (2007) argued that invariance across "all four measurement-elements is a necessary condition for MI [measurement invariance]" (p. 4). Thus, in the current study the proposed model will be examined across all levels of invariance (configural, weak, strong, and strict).

Current Study

As intelligence is thought to be an enduring trait, tests that measure intelligence should produce similar factor structures over time (Horn & McArdle, 1992). A cross-sectional analysis of the WISC-IV supported this assumption (Keith, et al., 2006). Unfortunately, cross-sectional analyses may not be adequate for detecting change over time (Willett, Singer, & Martin, 1998). There is no evidence regarding the stability of the WISC-IV structure across time for the *same* individuals. If the structure changes over time then WISC-IV test-retest score differences cannot be unequivocally interpreted as reflecting changes in the underlying constructs, thereby limiting the construct validity of the WISC-IV as well as the appropriateness of using this measure to identify disabilities in children. Therefore, this study will use confirmatory factor analysis techniques to examine the factor structure of the WISC-IV across time for a clinically referred sample. It is hypothesized that configural, weak, strong, and strict factorial invariance of WISC-IV scores across time will be demonstrated for a clinically referred sample.

Chapter 2

METHOD

Participants

Three hundred and fifty-two students who were twice administered the WISC-IV, with all ten core subtests administered at each test session, served as participants in the current study. Participant ages ranged from 6.1 to 14.11 years, with approximately 66% males (*n* = 231) and 34% females (*n* = 121). Of these participants, 3.1% were in first grade, 30.4% in second grade, 19.9% in third grader, 13.4% in fourth grade, 8% in fifth grade, 4.8% in sixth grade, 1.7% in seventh and eighth grades, and 0.6% in ninth through twelfth grades at first testing. Reported ethnic breakdown of the sample was 79% White, 11% Hispanic, and 6% Black, with 97% of students' primary home language being English. Approximately 95% of the students were eligible for special education services based on their primary diagnosis: 64.5% with learning disabilities, 12.5% with other health impairments (including attention deficit hyperactivity disorder), 7.5% with emotional disabilities, 4.6% with autism spectrum disorder (including Asperger's disorder), 2.6% with mental retardation, 2.3% with speech and language impairments, and 1.0% with other disabilities (hearing impairment and multiple disabilities).

Instrument

The Wechsler Intelligence Scale for Children-Fourth Edition (WISC-IV) is an individually administered intelligence test used for children between the ages of 6 and 16 years. The WISC-IV is a revised edition of the WISC-III and has been standardized with a nationally representative sample of 2,200 children ages 6 through 16 years. According to the technical manual (Wechsler, 2003b), the WISC-IV normative sample was representative of the U.S. population of children aged 6 to 16 years (March 2000 Census). Using a process of stratified sampling, the final normative sample consisted of 2,200 children with 100 boys and 100 girls in each one-year age group. The sample was stratified according to race/ethnicity and parent education level and geographic region (four major regions: Northeast, South, Midwest, and West). Exclusionary criteria for the standardization sample included previous intelligence testing within the past 6 months, uncorrected visual impairment, uncorrected hearing loss, non-English fluency, nonverbal/uncommunicative, disability that affects upper extremity motor performance, current hospitalization (medical, mental or psychiatric), current use of medication that may depress performance (antidepressants, anticonvulsants, antipsychotics etc.), and diagnosis of physical condition/illness that may depress testing performance (stroke, epilepsy, traumatic brain injury, meningitis, etc.). Approximately 5.7% of the normative sample consisted of exceptional children (children with disabilities or giftedness) in accordance to the current population of school age children at the time of test development.

The WISC-IV consists of 15 subtests, 10 core and 5 supplemental, each with a mean of 10 with a standard deviation of 3. The core subtests are Block Design, Similarities, Digit Span, Picture Concepts, Coding, Vocabulary, Letter-Number Sequencing, Matrix Reasoning, Comprehension, and Symbol Search. The supplemental subtests include Picture Completion, Cancellation, Information, Arithmetic, and Word Reasoning. The 10 core subtests are used to form four factor indices: Verbal Comprehension Index (VCI; Similarities, Vocabulary, and Comprehension), Perceptual Reasoning Index (PRI; Block Design, Matrix Reasoning, and Picture Concepts), Working Memory Index (WMI; Digit Span and Letter-Number Sequencing), and Processing Speed Index (PSI; Coding and Symbol Search). Each index score has a mean of 100 with a standard deviation of 15. A Full Scale Intelligence Quotient (FSIQ; *M* = 100; *SD* = 15) can also be formed from the 10 core subtests.

The technical manual for the WISC-IV reported strong internal consistency reliability coefficients in both the standardization and the special education samples (Wechsler, 2003b). For example, within the standardization sample the internal consistency coefficients for the test’s four indices are as follows: VCI = .94, PRI = .92, WMI = .92, PSI = .88 and FSIQ = .97. Reliability coefficients for the special education sample were not reported; however, internal consistency reliability coefficients were reported for eight of the ten core subtests within the special education sample (coding and symbol search were not used in this analysis). These coefficients ranged from 0.87 (Digit Span) to 0.93 (Letter-Number Sequencing and Matrix Reasoning). Additionally, Wechsler (2003b) reported strong correlation coefficients between the WISC-IV and other Wechsler scales including the WISC-III, the Wechsler Primary and Preschool Scale of Intelligence-Third Edition (WPPSI-III, Wechsler, 2002), the Wechsler Adult Intelligence Scale-Third Edition (WAIS-III; Wechsler, 1997), and the Wechsler

Abbreviated Scale of Intelligence (WASI; Wechsler, 1999), which are indicative of robust convergent validity.

Exploratory and confirmatory factor analyses indicated evidence of structural validity for the normative sample. Specifically, the technical manual reported that the first-order four-factor oblique structure was the best fit for the core subtests (Wechsler, 2003b). This structure has also been replicated by independent research for the normative sample (Keith et al., 2006; Watkins, 2006) and clinical samples (Bodin et al., 2009; Watkins, 2010; Watkins et al., 2006). Independent studies have also further assessed the normative sample and the structure of the WISC-IV by evaluating a multi-level structure. Independent research has indicated that the direct (Keith, 2005; Watkins, 2010) and indirect hierarchical (Bodin et al., 2009) models are the best fitting models. Additionally, studies have replicated that the general intelligence factor explains more variance than any of the first order factors (Watkins, 2006; Watkins, et al., 2006; Bodin et al., 2009; Watkins, 2010).

Procedure

Following IRB and school district approval, seven doctoral school psychology students reviewed special education files and extracted relevant WISC-IV data from approximately 7,500 student files in two participating school districts. The two participating school districts encompass forty-seven elementary schools, fourteen middle schools, three K-8 schools, eleven high schools, and one alternative school (K-12). One district serves 33,500 students and the second serves 26,600 students. School district demographics were collected from information provided by the National Center for Education Statistics (http://nces.ed.gov/). The first district is comprised of approximately 84% non-Hispanic or Latino students, with 6% of their students identified as English Language Learners. The second district is comprised of approximately 88% non-Hispanic or Latino students, with 4% of their students' identified as English Language learners.

Special education files were reviewed individually to determine if a WISC-IV was administered to each student. Approximately 3,111 files met this criterion with approximately 66% male (*n* = 2,059) and 34% female (*n* = 1,052). This initial sample consisted of 10.6% first graders, 15.8% second graders, 13.4% third graders, 14.2% fourth graders, 13.3% fifth graders, 12.5% sixth graders, 12.5% seventh and eighth graders, and 5.7% ninth thru twelfth graders. The ethnic composition of this sample was 76% White, 13% Hispanic, 5% Black, and 2% American Indian students with 94% of students' primary home language being English. Approximately 92% of the students were eligible for special education services based on their primary diagnosis: 57.2% with learning disabilities, 12.3% with other health impairments (including attention deficit hyperactivity disorder), 11.9% with emotional disabilities, 3.8% with autism spectrum disorder (including Asperger's disorder), 2.6% with mental retardation, 2.3% with speech and language impairments, and 1.0% with other disabilities (hearing impairment and multiple disabilities). Of those 3,111 files, 352 contained a second WISC-IV that met all selection criteria. Those 352 students served as participants in the current study (see Tables 2 and 3 for individual characteristics of the overall and test-retest sample).

Table 2

Grade Level for Overall and Test-Retest Sample

Grade	Overall Sample		Test-Retest Time 1		Test-Retest Time 2	
	Frequency	Percent of Sample	Frequency	Percent of Sample	Frequency	Percent of Sample
1	330	10.6	11	3.1	4	1.1
2	492	15.8	107	30.4	5	1.4
3	418	13.4	70	19.9	23	6.5
4	441	14.2	47	13.4	62	17.6
5	414	13.3	28	8.0	100	28.4
6	390	12.5	17	4.8	82	23.3
7	196	6.3	5	1.4	29	8.2
8	194	6.2	1	0.3	23	6.5
9 - 12	177	5.7	2	0.6	19	5.4

Table 3.

Individual's Characteristics for Overall and Test-Retest Sample

Characteristic	Overall Sample		Test-Retest Sample	
	Frequency	Percent of Sample	Frequency	Percent of Sample
Ethnicity				
American Indian	53	1.7	3	0.9
Asian/Pacific	47	1.5	4	1.1
Black	169	5.4	22	6.3
Hispanic	411	13.2	38	10.8
White	2354	75.7	279	79.3
Other/Missing Data	77	4.2	0	0
Special Education Eligibility				
Learning Disability	1779	57.2	227	64.5
Other Health Impairment	383	12.3	44	12.5
Emotional Disability	371	11.9	27	7.5
Not Eligible	261	8.4	18	5.1
Autism Spectrum	116	3.8	16	4.6
Mental Retardation	81	2.6	8	2.3
Speech and Language	73	2.3	8	2.3
Hearing Impaired	19	0.6	1	0.3
Multiple Disabilities	13	0.4	2	0.6
Primary Language				
English	2934	94.3	343	97.4
Spanish	93	3.0	4	1.2
Other	84	2.7	5	1.4

Statistical Analyses

Model specification. There are a number of analyses that can be applied to determine if the WISC-IV is measuring the same constructs with the same accuracy across time. Modern approaches evaluate invariance through multiple-group CFA (Byrne, 2006). However, when testing longitudinal invariance there is no categorization of multiple groups or samples; thus, the testing of longitudinal invariance requires that separate models at each time be fit simultaneously to the data (Wang, Elhai, Dai, & Yao, 2010). In accordance with previous empirical work and intelligence theory, three alternative models have been identified as best fitting and most appropriate in the normative and clinical samples: direct-hierarchical model, correlated four-factor model, and the indirect-hierarchical model. However, for the purpose of the current study, the direct-hierarchical model will be excluded because it will not be statistically identified without a constraint of equality of factor loadings, which will not allow for subsequent invariance tests. Thus, the remaining two models will be used for the first stage of analysis. Each model will be evaluated to determine the baseline model for the current analysis. A baseline model will be identified at both test and retest and the fit of each model (indirect-hierarchical and correlated four-factor) will be determined. Fit of the models will be compared to one another, and the best fitting model will then be used as the baseline model for further examination of factorial invariance.

For the identified baseline model factorial invariance testing will ensue. Initially, configural invariance will be assessed (see Figure 7 for an example of the indirect hierarchical model and Figure 8 for an example of the correlated four-factor model). Test and retest factor models will be constructed with factor correlations estimated between data from the first administration and data from the second administration. In the hierarchical model, the pairs of disturbance variances of the first-order factors at test and retest will be correlated; whereas, in the correlated four-factor model the pairs of factors will be correlated at test and retest among the four factors. Additionally, all pairs of residual error variances of the subtests will be correlated at test and retest because the same items were used to create the subtest score at both time points; more specifically, a subtest's residual error variance at test will be permitted to covary with that subtest's residual error variance at retest. The metric will be set at one factor loading for each first-order factor and in the hierarchical model the metric will be set at one factor loading of the second-order factor as well as one factor loading for each first-order factor. If invariance is found at this level it posits that the overall factor structure is the same at test and retest (configural invariance).

Upon confirmation of configural invariance, further constraints will be imposed on the model and compared to the configural invariance model. Thus, the fit of the proposed weak factorial invariance model will be assessed. To do this, the configural invariance model will be used and the remaining factor loadings will all be constrained to be equal across administrations. If the baseline model is a hierarchical model, then initially the first-order factor loadings will be constrained to be equal across time and assessed for invariance followed by the second-order factor loadings will be constrained to be equal across administrations (Chen et al., 2005). A chi-square difference test will be used to assess if the constrained model (weak factorial invariance) is significantly different from the baseline model (configural invariance). If the chi-square difference test is not significant then it is indicative that the magnitude of the factor loadings are the same across test and retest, satisfying weak invariance.

Next, the fit of the proposed strong factorial invariance model will be evaluated by using the weak factorial invariance model and constraining the intercepts to be equal across administrations. If the baseline model is a hierarchical model, then initially the intercepts of measured variables (subtests) will be constrained to be equal across time, followed by the intercepts of the first-order factors (Chen et al., 2005). If invariance is found at this level, it indicates that the test is measuring the same construct across time with similar accuracy.

Finally, assuming strong factorial invariance is upheld, the fit of the proposed strict factorial invariance model will be tested. The strong factorial invariance model will be used and constraints of equal subtest error variances will be applied across administrations. If the baseline model is a hierarchical model, then initially the first-order disturbances will be constrained to be equal across time followed by the residual variances of the observed variables (Chen et al., 2005). This sequence of invariance tests will be conducted in a similar manner to Grouzet, Otis, and Pelletier (2006) and Wang et al. (2010). Confirmatory factor analysis as implemented in Mplus 6.11 (Muthén & Muthén, 2010) will be used to test these sequential levels of factorial invariance.

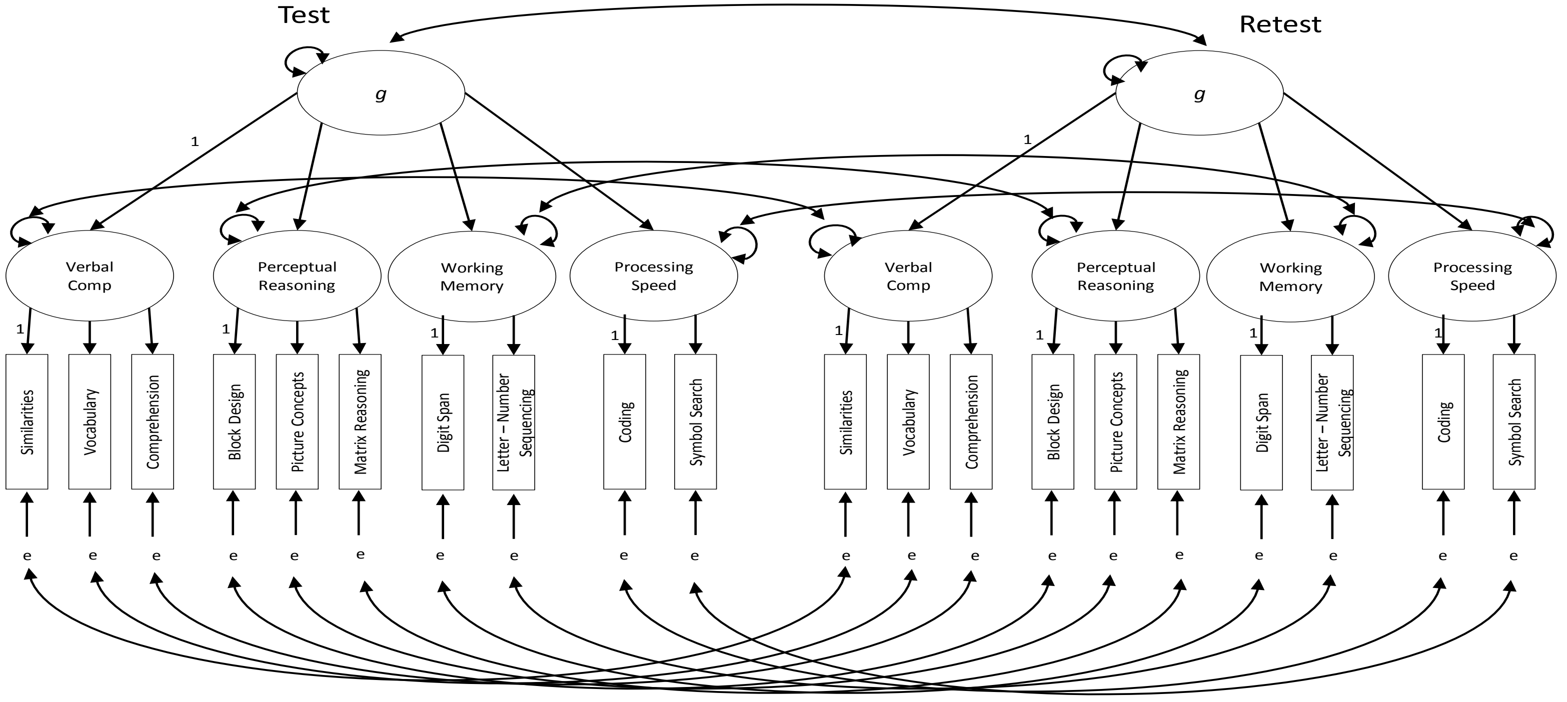


Figure 7. Assessing configural invariance of the proposed indirect hierarchical model at Time 1 and Time 2.

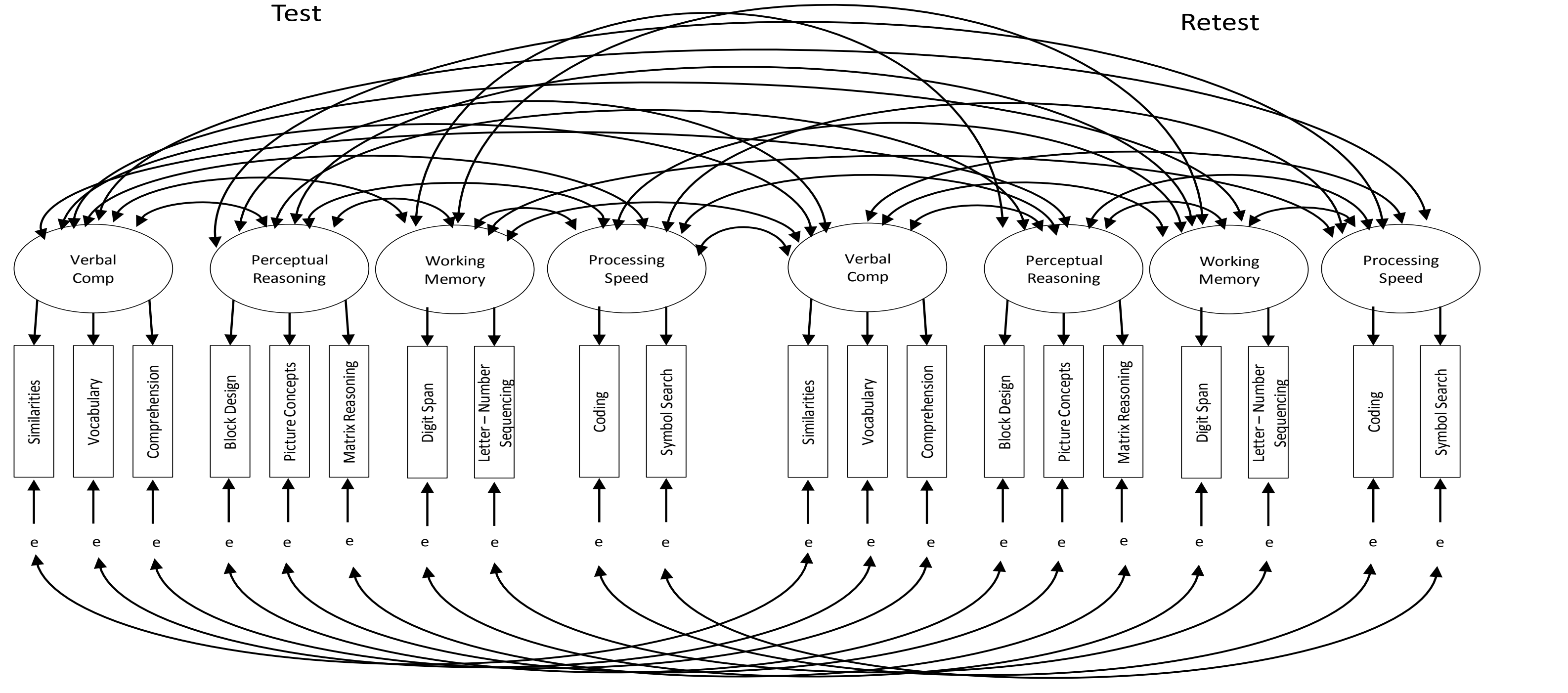


Figure 8. Assessing configural invariance of the proposed correlated four-factor model at Time 1 and Time 2.

Comparison of fit between models. For the purpose of this analysis, fit indices and model comparison statistics will be reported. The following indices will be used to determine fit of each model: the chi-square statistic (χ^2); the comparative fit index (CFI; Bentler, 1990), root-mean-square-error of approximation (RMSEA), Bayesian Information Criterion (BIC; Raftery, 1995), and the standardized root mean square residual (SRMR). Fit will be determined to be good, acceptable, or unacceptable. The χ^2 index is considered to be a "badness of fit" index because the model's fit is worse when the χ^2 value is high. The CFI is an index used to evaluate model fit. This index compares the "null" model, where all measured variables are uncorrelated, to the model being tested. Unlike the χ^2 index, the CFI does not assume perfect population fit and appears to be robust in large and small samples (Keith, 2005). An additional fit index that is used is the Bayes Information Criteria (BIC; Raftery, 1995). This fit index penalizes complex models by taking into account the number of free parameters. A lower BIC value indicates a better fitting model. SRMR is another fit index that is commonly used; this index represents the overall difference in correlations between the observed and predicted models (Keith, 2005). In accordance with Hu and Bentler (1999), a value of $\geq .95$ on the CFI, a value of $\leq .06$ for the RMSEA, and a value of $\leq .08$ for the SRMR will be used to indicate good fit between the theoretical model and the sample data. Acceptable fit will be indicated by a CFI value of $\geq .90$ (Hu & Bentler, 1999), a RMSEA value $\leq .08$ (Browne & Cudeck, 1993), and a SRMR value $\leq .10$ (Kline, 2005). Values of CFI $< .90$, RMSEA $> .08$, and SRMR $> .10$ will signal unacceptable fit.

The χ^2 difference test (Satorra & Bentler, 2001) will be used to assess differences between models. A significant χ^2 difference test is indicative of differences in the fit of the two models and thus would not lend evidence of good fit whereas a non-significant χ^2 difference test would indicate invariance across test and retest. In order to account for potential experimentwise errors caused by using multiple chi-square difference tests, a Bonferroni correction was applied to the alpha level (*p*). An initial overall significance level of $p < .05$ was used but due to multiple tests the Bonferroni correction was applied so that individual tests were at the $p < .0125$ (.05/4) level (Green, Thompson, & Babyak, 2010). As the χ^2 difference test is sensitive to sample size (Keith, 2005), additional tests will be used to establish evidence of factor invariance. Cheung and Rensvold (2002) identified that an alternative to using the χ^2 difference test is by evaluating the change (Δ) in another general fit index. Based upon results of a simulation study, it was determined that measuring change in the comparative fit index (ΔCFI) is a robust statistic for testing between-group invariance of CFA models. This study determined that a ΔCFI value of ≤ 0.01 is indicative of invariance. Chen (2007) also supported this cutoff value for evaluating fit. CFI difference values between .01 and .02 are indicative of mean differences and CFI differences >.02 indicate definite differences (Cheung & Rensvold, 1999, 2002). Additionally, the Bayesian information criterion (BIC) difference test will be used to test models that are not nested. A difference of ≥ 10 is indicative of very strong support, 6-10 points is indicative of strong support, 2-6 is indicative of positive support, and < 2 is indicative of weak support, for the model with the lower BIC value (Raftery, 1995).

Chapter 3

RESULTS

Descriptive statistics for WISC-IV subtest, factor, and IQ scores at test and retest for the referred special education sample are reported in Table 4. These results indicate that the current sample exhibited slightly lower and more variable scores than the normative sample of the WISC-IV (Wechsler, 2003b). This pattern of scores has been observed in similar samples of students referred for special education evaluations (Watkins et al., 2006). The univariate score distributions from the current sample appear to be relatively normal across both test administrations, with .43 the largest skew and .93 the largest kurtosis at test as well as -.48 the largest skew and .91 the largest kurtosis at retest. Additionally, examination of each variable's associated histogram indicated that the sample appears to generally follow the shape of a normal distribution (Tabachnick & Fidell, 2007). Although the univariate skewness and kurtosis statistics indicated normality, subsequent analyses require multivariate normality. The *Mplus* program provided multivariate skewness (5.71 and 5.83) and kurtosis (124.32 and 126.49) statistics based on Mardia (1970) for test and retest occasions, respectively. However, Muthén (2011) indicated that tests of multivariate normality are no longer as important as in the past because there are now non-normality robust techniques that can be applied. Consequently, the MLM robust estimation technique was used, rather than ML estimation, to adjust for non-normality.

Table 4.

Mean, Standard Deviations, Skewness, and Kurtosis of Wechsler Intelligence Scale for Children-Fourth Edition (WISC-IV) Subtest, Factor, and IQ Scores of 352 Students Twice-Tested for Special Education Eligibility

Variable	Mean		*SD*		Skewness		Kurtosis	
	Test	Retest	Test	Retest	Test	Retest	Test	Retest
BD	9.2	8.7	2.8	3.0	-0.07	0.24	0.05	-0.08
SI	8.8	6.2	2.6	2.8	0.10	0.06	0.14	0.29
DS	8.0	7.8	2.6	2.6	0.09	-0.22	0.93	0.07
PCn	9.5	10.0	3.3	3.0	-0.20	-0.48	-0.07	0.35
CD	8.4	7.5	3.2	2.9	0.43	0.00	0.05	-0.15
VC	8.6	8.4	2.7	2.7	0.09	-0.07	0.17	0.14
LN	8.1	8.2	2.8	3.1	-0.37	-0.73	-0.11	-0.10
MR	9.1	9.1	3.0	3.1	0.19	0.01	0.33	0.16
CO	8.9	8.9	2.7	2.6	-0.24	-0.60	0.66	0.91
SS	8.4	8.7	3.3	3.1	-0.28	-0.18	0.02	0.26
VCI	92.5	93.0	12.7	13.2	-0.21	-0.20	0.67	0.81
PRI	95.5	95.4	15.0	15.7	-0.33	-0.28	0.20	0.25
WMI	88.3	88.0	13.0	14.2	-0.23	-0.56	-0.67	0.23
PSI	91.3	89.3	15.1	15.0	0.03	0.05	-0.08	0.10
FSIQ	90.3	89.9	13.6	14.5	-0.39	-0.40	0.67	0.82

Note. BD = Block Design; SI = Similarities; DS = Digit Span; PCn = Picture Concepts; CD = Coding; VC = Vocabulary; LN = Letter-Number Sequencing; MR = Matrix Reasoning; CO = Comprehension; SS = Symbol Search; VCI = Verbal Comprehension Index; PRI = Perceptual Reasoning Index; WMI = Working Memory Index; PSI = Processing Speed Index; FSIQ = Full-Scale IQ.

Baseline Model Identification

Previous research indicates that the correlated four-factor, direct-, and indirect-hierarchical models have been identified as the best fitting and most appropriate models for the WISC-IV (Keith, 2005; Watkins et al., 2006; Gignac, 2008; Bodin et al., 2009) . However, the direct-hierarchical model was excluded due to a failure to achieve statistical identification. The remaining models were evaluated to determine the best fitting baseline model for the current study. Each model was evaluated for fit at test and retest (See Table 5 for goodness-of-fit indices for both models at test and retest). According to the goodness of fit indices, both models indicated relatively good fit within each individual time point (both test and retest).

Table 5.

Goodness-of-fit indices for baseline models at test and retest.

Model	χ^2	*df*	CFI	BIC	RMSEA	SRMR
Correlated Four-Factor model at Test	61.8	29	.97	16367.74	.06	.03
Correlated Four-Factor model at Retest	93.0	29	.96	16087.08	.08	.04
Indirect Hierarchical Model at Test	62.6	31	.97	16362.9	.05	.03
Indirect Hierarchical Model at Retest	100.1	31	.95	16089.05	.08	.05

The chi-square difference test identified that at test there was no significant change in fit between the models; however, at retest there was a significant difference according to the chi-square difference test. Since the indirect-hierarchical model is nested within the correlated four-factor model, the correlated four-factor model was chosen as the baseline model as it was the supported model at retest based on the chi-square difference testing (See Table 6 for model fit comparison statistics at test and retest).

Table 6.

Model fit comparison statistics at test and retest.

Model Comparisons	$\Delta\chi^2$	Δ*df*	*p* value	ΔCFI	ΔBIC
Indirect and Correlated models at Test	.701	2	.70	.001	-4.85
Indirect and Correlated models at Retest	7.063	2	.03	.003	1.97

Factorial Invariance

The correlated four-factor model was used as the baseline model for invariance testing. Invariance testing was conducted using *Mplus* 6.11 (Muthén & Muthén, 2010) and followed the sequence of invariance tests described by Wang et al. (2010). Invariance testing was conducted across configural, weak, strong, and strict levels of testing for the correlated four-factor model. Each level of invariance was achieved prior to continuation of invariance testing (See Table 7 for all invariance testing results). All chi-square difference tests were conducted with the modified formula described by Muthén & Muthén (2010) to account for robust ML methods.

Configural invariance. The correlated four-factor model identified as the baseline model was tested initially for configural invariance. Upon initial investigation *Mplus* identified a possible linear dependency among the latent variables associated with working memory at test and retest (WMI1 and WMI2). This linear dependency indicated high correlation across time. To correct for this dependency, the covariance between the offending parameters (WMI1 and WMI2) was fixed to equal 1. The raw data was used to check the correlation between WMI1 and WMI2 (*r* = .65) and this correlation was compared to the correlation between WMI1 and WMI2 with the constraint in place (*r* = .69). As these correlations were similar, the constraint was left in place for the remainder of invariance testing. Continuation of configural invariance testing resulted in good fit (χ^2 = 255.46 (133), CFI = .965, RMSEA = .051 (.042 - .061), SRMR =

.083, BIC = 31655.78). These results indicate that configural invariance was upheld across time and thus the sequence of invariance testing can continue.

Factor loadings. Standardized factor loadings for the correlated four-factor model at the level of configural invariance testing are presented in the path diagram in Figure 9. All factor loadings of the observed variables were moderate to high and significant (p < .001) ranging from .50 to .92. The standardized factor loadings for the correlations amongst the latent variables were also moderate to high and significant (p < .001) ranging from .31 to .92. These results indicate that all observed variables loaded appropriately on the indicated factors at both test and retest.

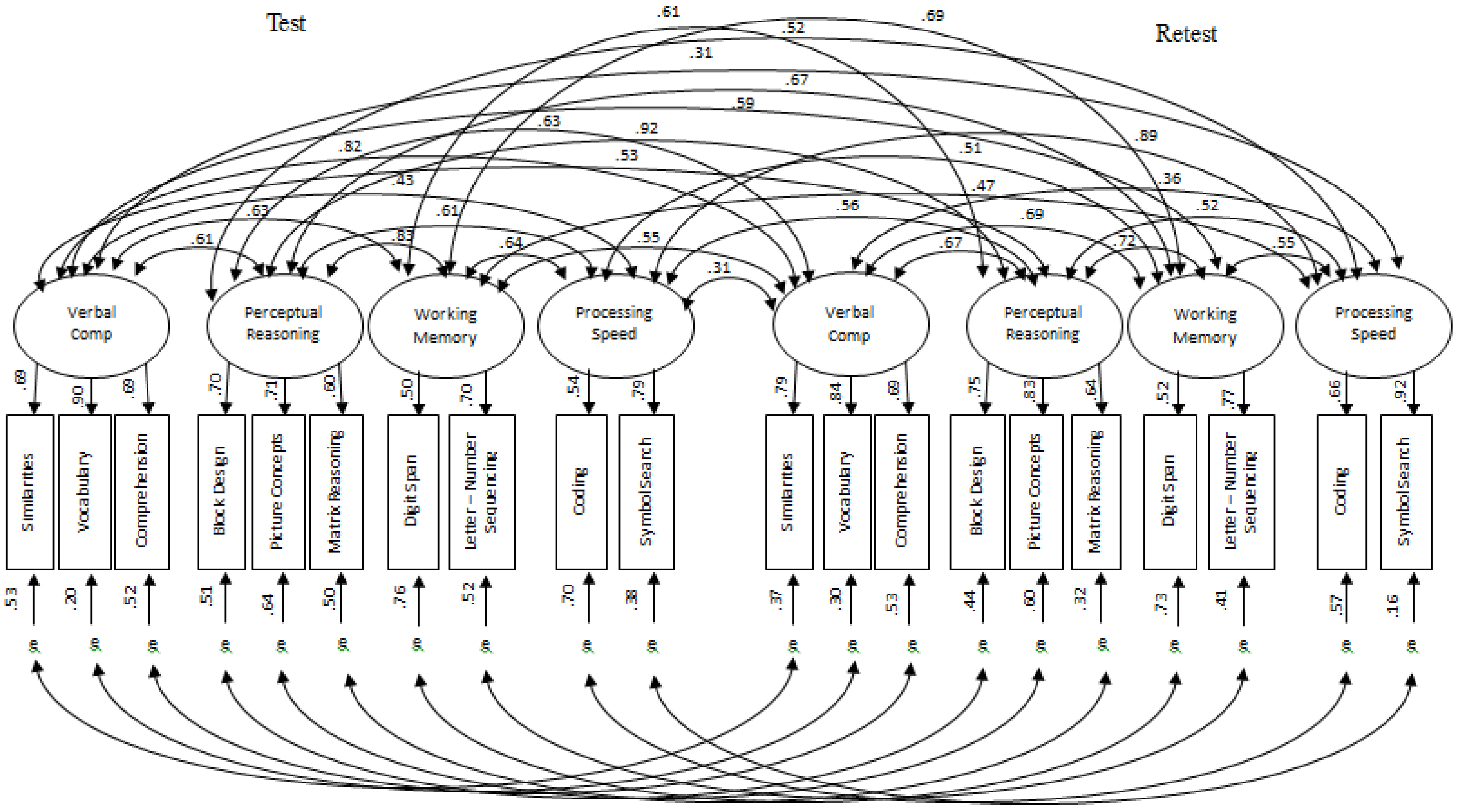


Figure 9. Path diagram of correlated four-factor model at test and retest

Note. All loadings are significant at $p < .001$, except for the error term associated with Symbol Search 2, $p = .06$.

Weak invariance. For weak factorial invariance testing, the factor loadings were constrained to be equal across time for the configural model. This level of invariance testing indicated good fit overall (χ^2 = 270.59 (139), CFI = .962, RMSEA = .052 (.043 - .061), SRMR = .085, BIC = 31655.337). The chi-square difference test was not significant, meaning that there was no statistically significant change between the configural and weak invariance models at the a priori specified significance level, *p* = .0125, ($\Delta\chi^2$ = 15.06 (6), *p* = .02). Weak invariance was also supported by a CFI difference less than .01 and a BIC difference less than 2 (ΔCFI = .003, ΔBIC = .443). To ensure that no individual indicator would significantly change the fit of the overall model, each pair of observed variables was unconstrained one at a time. This check resulted in minor to no change in fit and thus all pair of indicators remained constrained to be equivalent across time.

Strong invariance. Next, the observed variable intercepts were constrained to be equal across time. Initial results indicated acceptable fit overall (χ^2 = 349.947 (149), CFI = .942, RMSEA = .062 (.054 - .07), SRMR = .087, BIC = 31709.14). The chi-square difference test was statistically significant, meaning that there was change between the weak and strong invariance models ($\Delta\chi^2$ = 80.18 (10), *p* < .001). Additionally, the CFI difference was greater than .01 and the BIC difference was greater than 10 (ΔCFI = .02, ΔBIC = 53.803). As these results indicate significant change between the models, there is no support for full strong invariance. In order to evaluate for partial invariance at this stage, recommended modifications were obtained from *Mplus* on several observed variable intercepts. The affected variables were the Coding subtest (CD1 and CD2), Block Design subtest (BD1 and BD2), and Similarities subtest (SI1 and SI2).

First, the previously constrained intercepts of CD1 and CD2 were released. This resulted in improved fit of the model, (χ^2 = 317.676 (148), CFI = .951, RMSEA = .057 (.048 - .066), SRMR = .086, BIC = 31678.879). However, the chi-square difference test comparing the weak invariance model to the current more constrained model ($\Delta\chi^2$ = 47.54 (9), p < .001) remained significant, the difference in CFI was above .01 and the difference in BIC was above 10 (ΔCFI = .011, ΔBIC = 23.342), indicating that strong invariance continued to fail to be upheld.

Second, in addition to the released CD1 and CD2 constraints, the previously constrained intercepts between BD1 and BD2 were released. This model resulted in improved fit (χ^2 = 297.08 (147), CFI = .957, RMSEA = .054 (.045- .063), SRMR = .085, BIC = 31660.549). As, the chi-square difference test comparing the weak invariance model with the current more constrained model remained significant ($\Delta\chi^2$ = 26.68 (8), p < .001), partial strong invariance was not upheld due to the change in BIC indicating positive support of change (ΔCFI = .005, ΔBIC = 5.212).

Lastly, in addition to the released CD1, CD2, BDI, and BD2 constraints, the previously constrained intercepts between SI1 and SI2 were released. This model resulted in improved fit (χ^2 = 285.53 (146), CFI = .96, RMSEA = .052 (.043 - .061), SRMR = .085, BIC = 31651.473). The chi-square difference test comparing the weak invariance model with the current more constrained model was not significant ($\Delta\chi^2 = 14.97$ (7), $p = .04$) indicating support of partial strong invariance; additionally, the change in CFI indicated support of partial strong invariance (ΔCFI = .002). Although the change in BIC was above 2, indicating change in models (ΔBIC = 3.86), partial strong invariance was indicated in the model due to the non-significant results of the chi-square difference test and the small amount of change in the CFI value between models when the intercepts of the variables BD, CD, and SI were unconstrained at test and retest. Therefore, the WISC-IV factor loadings and factor intercepts (with the exception of CD, BD, and SI) were equivalent across test and retest. Thus, the WISC-IV exhibited both configural and weak invariance across time as well as partial strong invariance across time (with the exception of the CD, BD, and SI variables).

Strict invariance. The final step of invariance testing included constraint of the error variances associated with the observed variables to be equal across time. Due to the constraints removed from the CD, BD, and SI variables in the previous step, testing of strict invariance allowed those variables to remain unconstrained and new constraints were only applied to the remaining observed variables' error variances. Results indicated good fit overall (χ^2 =308.85 (153), CFI = .955, RMSEA = .054 (.045 - .062), SRMR = .086, BIC = 31657.847). The chi-square difference test was statistically significant ($\Delta\chi^2 = 22.8$ (7), $p = .002$); additionally, the change in BIC did not support invariance at this level. Although the CFI difference was less than .01, partial strict invariance was not upheld.

Additional modifications were recommended by *Mplus* on the error variance associated with the subtest, Symbol Search (SS); thus the previously constrained error variances of SS1 and SS2 were released. Upon release of this constraint, partial strict invariance was achieved. Results indicated good fit overall (χ^2 =300.64 (152), CFI = .957, RMSEA = .053 (.045 - .062), SRMR = .086, BIC = 31651.58). The chi-square difference test was not statistically significant ($\Delta\chi^2$ = 14.89 (6), *p* = .02); additionally, the change in BIC (-0.11) and the difference in CFI (.003) indicate that partial strict invariance is upheld. Therefore, the WISC-IV factor loadings, factor intercepts (with the exception of CD, BD, and SI), and error variances (with the exception of CD, BD, SI, and SS) were equivalent across test and retest. Thus, the WISC-IV exhibited configural, weak, partial strong (with the exception of the CD, BD, and SI variables), and partial strict invariance across time (with the exception of the CD, BD, SI, and SS variables).

Table 7.

Invariance Analysis Results for the Correlated Four-Factor Model Across Test and Retest.

Invariance Model	*df*	χ^2	CFI	RMSEA (C.I.)	SRMR	BIC	Δ *df*	Δ χ^2	Δ CFI	ΔBIC
1. Conf. Invariance	133	255.46	.965	.051 (.042-.061)	.083	31655.78	-	-	-	-
2. Weak Invariance	139	270.59	.962	.052 (.043-.061)	.085	31655.34	6	15.06	.003	.443
3. Strong Invariance	146	285.53	.96	.052 (.043-.061)	.085	31651.47	7	14.97	.002	3.86
4. Strict Invariance	152	300.64	.957	.053 (.045-.062)	.086	31651.58	6	14.89	.003	-0.11

Note. The results for strong invariance presented here are for the model with the intercepts of CD1, BD1, SI1, CD2, BD2, and SI2 unconstrained. The results for strict invariance presented here are for the model with residual error variances of

CD1, BD1, SI1, SS1, CD2, BD2, SI2 and SS2 unconstrained. If the decrease in the CFI value is .01 or greater, then the global test of the invariance constraints at the particular step does not hold. If the change in BIC value is 2 or greater, then the global test of the invariance constraints at that particular step does not hold.

Summary of Findings

Overall, the correlated four-factor model was the best fitting model at test and retest. Using the correlated four-factor model as a baseline, factorial invariance testing ensued. Configural and weak invariance were achieved, signifying that the overall factor structure remained the same at test and retest with equal precision of the factor loadings at both time points. However, strong invariance was not found; partial strong invariance was achieved by freeing the intercepts associated with the Block Design, Coding, and Similarities subtests. Finally, partial strict invariance was not obtained even when the errors associated with the Block Design, Coding, and Similarities subtests were freed and thus partial strict invariance was not achieved. Additional release of the error associated with the Symbol Search subtest resulted in achievement of partial strict invariance across time.

Chapter 4

DISCUSSION

The goal of the current study was to investigate factorial invariance of the WISC-IV for a group of 352 students eligible for psychoeducational evaluations tested, on average, 2.8 years apart. One research question was addressed in this study: Does the structure of the WISC-IV remain invariant for the same individuals across time in a referred sample? It was hypothesized that the factor structure of the WISC-IV would remain invariant, across all levels of invariance, in the same individuals across time with this referred sample. Using structural equation modeling methods this study found invariance across the configural and weak invariance levels and partial invariance at the strong and strict levels of invariance. Three subtest intercepts (BD, CD, and SI) were not equivalent across test and retest; additionally, four subtest error variances (BD, CD, SI, and SS) were not equivalent across test and retest. These results indicate that the WISC-IV measures the same constructs equally well across time.

Factorial Invariance

The identified baseline model, a correlated four-factor model, exhibited good fit and resulted in similar factor loadings as previous research of the WISC-IV suggested at both time points (Wechsler, 2003b; Watkins, et al. 2006). More importantly, the pattern of factor loadings remained similar across each model at test and retest. The factor loadings at both testing occasions indicated that each individual subtest had moderate to strong factor loadings on each assigned factor. This indicates that the individual subtests are in fact measuring each identified factor similar to the expected factor structure articulated by the author of the WISC-IV. Additionally, the pattern of correlations between latent variables indicated that the four factors are highly correlated with one another, providing evidence that there is likely a higher-order factor associated with the construct of intelligence. Furthermore, the pattern of correlations between latent variables remained similar across test and retest as well. As each of the factors was correlated with one another at each time point, the same pattern was also allowed across time; meaning that all factors at initial testing were allowed to correlate with all factors at the retesting period. The correlations between factors across time also followed a similar pattern to the patterns found at each testing time. This finding also lends evidence that there is a higher-order factor that is contributing to the correlations between these variables.

Configural invariance. The correlated four-factor model was tested for configural invariance and the results indicated that configural invariance was upheld across time. Verification of configural invariance indicates that each measured variable identically loads upon its specified common factor (Gregorich, 2006). Specifically, this requires that the same subtests are loading on each respective factor across time; meaning that the overall factor pattern is the same at test and retest. This indicates that the WISC-IV is measuring similar constructs at both test and retest occasions. Configural invariance is considered to be the least restrictive test of similarity of factors across time (Chen, 2007).

Weak invariance. The correlated four-factor model was further tested for weak factorial invariance with the results indicating that weak invariance was upheld. Because of the multiple significance tests, the Bonferroni correction was used to control the experimentwise alpha level, resulting in a significance level of .0125 for any single significance test. Due to this conservative significance level, weak invariance was achieved with the chi-square difference test. If the Bonferroni correction had not been utilized, this test would have been considered to be significant and thus weak invariance would not have been achieved. However, when completing factorial invariance testing, the chi-square difference test is frequently disregarded due to the chi-square's dependence upon sample size if the other model comparison statistics indicate invariance (Brannick, 1995; Kelloway, 1995; Wu, Li & Zumbo, 2007). Following this tradition, changes in CFI and BIC fit indices were not large enough to reject invariance. Thus, the conservative alpha level was not dispositive of weak invariance.

The achievement of weak invariance means that corresponding factor loadings are equivalent across groups (Gregorich, 2006). That is, each measured variable loads equivalently on each identified factor at both test and retest occasions. Thus, the constructs are being measured with equal precision at both occasions. This provides evidence that the identified factors of the WISC-IV (VC, PR, PS, and WM) have the same meaning across time. Therefore, it can be assumed that the factor structure of the WISC -IV remains stable across time and any interpretation of change in test scores can be unequivocally attributed to change in the constructs being measured and not to changes in the structure of the test itself. Specifically, this finding supports that average differences between test and retest factor scores can be compared. This means that the overall pattern of strengths and weaknesses identified by the WISC-IV can be compared for an individual across time points. For example, an individual who has a strength in Verbal Comprehension and a weakness in Processing Speed at initial testing should follow a similar pattern at retest. However, an individual's exact factor scores should not be directly compared and interpreted as change in an individual's cognitive ability between test and retest. In other words, practitioners should attend to the overall cognitive profile of the individual rather than focus on the specific factor scores.

A number of measurement researchers agree that achieving both configural and weak factorial invariance is enough evidence to determine that a measure is invariant across time, particularly in behavioral science research (Widaman & Reise, 1997; Horn, 1991; Bentler, 2005) and that further invariance testing is discretionary (Vandenbreg & Lance, 2000, Wu et al., 2007). Accordingly, this study continued to evaluate factorial invariance by addressing both strong and strict levels of invariance.

Strong Invariance. The correlated four-factor model was further tested for strong factorial invariance. Upon initial testing there was significant change between the models; thus, recommended modifications were obtained from *Mplus* on three observed variable intercepts. Partial strong invariance was achieved by releasing the previously constrained intercepts of the Coding, Block Design, and Similarities subtests. This indicates that for the majority of the subtests, scores have the same unit of measurement (factor loadings) as well as the same origin (intercept) and thus the factor means can be compared across time. Therefore, differences between the groups on the factor means demonstrates that there are true differences occurring across time and that it is not due to an artifact of the test itself (Chen et al., 2005). However, as the subtests (CD, BD, and SI) did not indicate invariance, this lends evidence that the factor means may not demonstrate true differences. Meaning that for the factors measured by the variant subtests (Processing Speed by CD, Perceptual Reasoning by BD, and Verbal Comprehension by SI) the means may not be interpreted as invariant because their constituent indicators did not remain stable across time.

The inability to achieve full strong invariance makes interpretation complicated. The data related to the subtests Coding, Block Design, and Similarities indicated that the factor means did not remain stable across time. This could indicate that these specific subtests are not as stable across time as the other subtests of the WISC-IV and thus any interpretation of change in these subtests across time should be done with caution. Horn (1991) indicated that achievement of configural and weak invariance "is a reasonable ideal for research in the behavioral sciences" (p. 124); verifying that the achievement of configural and weak invariance are necessary to support measurement invariance across time, but further invariance testing is not essential. In practical applications, it is typically appropriate to accept partial invariance as long as less than 20% of parameters are freed to achieve partial invariance (Dimitrov, 2010). Thus, it can be said that partial strong invariance was achieved as only 12% of parameters were freed to achieve partial invariance at this level. As full strong invariance is not a requirement of measurement invariance we can still say that the WISC-IV measures the same constructs equally well across time. Yet, due to a failure to achieve full strong invariance, it cannot be said with confidence that all factor and subtest means can be meaningfully compared across time. This finding indicates that the practice of comparing an individual's test scores at different time points should be completed with caution because, although one unit of change in tests scores can be considered to be equivalent to one unit of change in retest scores, the scales are not fully invariant.

Strict Invariance. Although strong invariance was not fully obtained, the partially invariant model was used to test for strict factorial invariance. The results indicated that partial strict invariance was not upheld. The same subtests that did not allow strong invariance did not allow strict invariance. Unfortunately, partial strict invariance was not achieved by releasing the error constraints of the same subtests as in the strong invariance testing (Coding, Block Design and Similarities) and there was need to release an additional error constraint. Recommended modifications were obtained by *Mplus* on one observed variable's error term. Partial strict invariance was thus achieved by releasing the previously constrained error associated with the Symbol Search subtest.

With the exception of the Coding, Block Design, Similarities, and Symbol Search subtests, differences between test and retest scores across time were due to group differences associated with the common factors. Thus, allowing the assumption that unique error variances are not contributing to differences in test scores across time. The data related to the subtests Coding, Block Design, Similarities, and Symbol Search indicated that the unique error variances did not remain stable across time. This could indicate that these specific subtests are not as stable across time as the other subtests of the WISC-IV and thus any interpretation of change in these subtests across time should be done with caution. Specifically, within these subtests it cannot be assumed that unique error variances are not contributing to differences in test scores across time.

As error variances of tests are not typically expected to be equal, the failure to obtain strict factorial invariance does not invalidate equivalence of the factor structure of the WISC-IV (Bentler, 2005; Byrne, 2012; Marsh, 1993; Watkins & Canivez, 2001). Although it is ideal to achieve more stringent levels of invariance, only configural and weak invariance are required to indicate invariance of a measure. Thus, these data continue to support the hypothesis that the WISC-IV measures the same constructs equally well across time.

Limitations

As with all research, there are a number of limitations in the current study that should be improved upon in future studies. The greatest of these limitations is the sample. Although typically a sample of 352 students is considered to be large, this is a relatively small sample for completing factorial invariance testing of complex structures. Ideally, a larger sample is desired when completing these types of analysis (Byrne, 2012). An additional limitation of this study is the method of data collection. As the data was collected from archived special education records, administration and recording accuracy of the individual psychologists who administered the WISC-IV had to be assumed. Moreover, although there was training for the graduate students who collected this data it is possible that data entry errors may have occurred.

The sample used in this study was from two school districts in central Arizona and thus may not be generalizable to other regions. This sample was largely identified as Caucasian, non-Hispanic students (80%). With a small percentage of students from a minority background, any current findings may not be generalizable to samples of students from other racial backgrounds. Additionally, the majority of the students in these school districts do not qualify for free and reduced lunch (94%); thus, results of this study may not be generalizable to different levels of socio-economic status. Finally, the sample consisted largely of English speaking students (94%); however, the available data did not include the English language proficiency of individual students and thus some students' results may have unknowingly been affected by their level of English proficiency even though the special education records indicated that they were English speakers. There is a large body of research that indicates variability of scores on cognitive assessments for students with limited English proficiency (Anastasi & Urbina, 1997; Frisby, 1999; Hays, 2008; Schon, Shaftel, & Markham, 2008). Thus, these results need to be further examined in regards to students that are not English language proficient.

Furthermore, the sample consisted solely of students referred for a psychoeducational evaluation for special education eligibility. The current sample appears to have an overrepresentation of students identified as children with specific learning disabilities as compared to the national average, making it a very selective sample.   Due to the specificity of the current sample, it is unclear if the same results would apply to other referred samples of students as well as non-referred samples of students. Students were also excluded from this sample if they had not been administered the WISC-IV more than once, causing multiple individuals to be excluded from the study who did not qualify for special education at the initial test administration.

Finally, the data sample did not consist of many WISC-IV test administrations with complete WISC-IV records. In other words, few records indicated supplemental test scores and therefore only core subtests could be used in the current analyses. Previous research indicates that clinicians are unlikely to administer supplemental subtests (Watkins et al., 2010); however, a limitation of this study is that only core subtests were used when examining the factor structure of the WISC-IV. If all subtests had been administered then the direct-hierarchical model would have been identifiable and could have been included when determining the best baseline model.

Implications for Practice. The scope of this study was not intended to provide concrete recommendations for clinicians utilizing the WISC-IV, and due to the previously discussed limitations it is not recommended that the findings of this study be generalized to the entire population or even to other selected samples. However, if the results of this study were to be replicated with multiple larger, more diverse samples that more appropriately matched the national population, then there are a few recommendations that could be made for clinician’s using the WISC-IV. The results of this study, along with replication studies, allows clinicians to be confident in interpretation of differences in an individual's overall cognitive profile pattern across time as a reflection of change in the constructs and not as change in the measure itself. More specifically, a change in a student's subtest or index scores across test administrations could indicate an actual change within the individual's ability and any such discrepancy should be followed up according to best practice or district policies.

Future Research

Additional research is needed to further understand the invariance of cognitive assessment scores across time. As this study was not intended to determine the correct factor model of the WISC-IV, but to examine longitudinal factor invariance, the results of this study may not generalize across alternative identified factor structures of the WISC-IV and should be further examined. Accordingly, future research should focus on alternative identified factor structures of the WISC-IV (specifically, the indirect- and direct- hierarchical models) to determine if the current results are generalized across factor structures. The results of the current study indicated that there are moderate to strong correlations between factors across time. This is evidence of a likely higher-order factor that can explain greater amounts of variance across time; it is essential that further evaluation occur. As the indirect-hierarchical model has been identified as the best model for previous versions of the WISC (Bodin et al., 2009), this factor structure is especially important to examine for invariance across time. Additionally, further research is needed with more diverse populations as well as non-exceptional children in order to determine the generalizability of the current results.

Longitudinal research studies need to continue in the area of cognitive assessment in order to better understand the stability of the latent constructs being measured. This is true for the WISC-IV, its successors, and other individual tests of intelligence. Without evidence to support that the constructs of individual assessments remain constant across time, practitioners cannot appropriately interpret results. Every time a new test is revised it is essential that further evidence be collected regarding the construct validity of the new version of the test and that it is not simply assumed that the constructs remain stable.

Conclusion

The Wechsler scales of intelligence are the most frequently used intelligence tests among clinicians (Alfonso, Oakland, LaRocca, & Spanakos, 2000; Belter & Piotrowski, 2001; Pfeiffer, Reddy, Kletzel, Schmelzer, & Boyer, 2000). The most recent version of this test for children is the Wechsler Intelligence Scale for Children - Fourth Edition. It is assumed by clinicians that the structure of the test remains invariant across time and that an individual’s scores are comparable across time; however, this has not previously been empirically verified. The current study is the first to investigate the factor structure of the WISC-IV across time for the same individuals. While the current study found configural and weak invariance, only partial invariance was found at the strong and strict levels of invariance and additional research is recommended. However, as only configural and weak levels of invariance are required for measurement invariance to be achieved, these data support the hypothesis that the WISC-IV measures the same constructs equally well across time.

REFERENCES

Alfonso, V. C., Oakland, T. D., LaRocca, R., & Spanakos, A. (2000). The course on individual cognitive assessment. *School Psychology Review, 29*, 52-64.

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

American Psychological Association. (2002). *Ethical principles of psychologists and code of conduct*. Washington, DC: Author.

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

Anyon, Y. (2009). Sociological theories of learning disabilities: Understanding racial disproportionality in special education. *Journal of Human Behavior in the Social Environment, 19*, 44-57.

Belter, R. W., & Piotrowski, C. (2001). Current status of doctoral-level training in psychological testing. *Journal of Clinical Psychology*, *57*, 717-726.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 197*, 238-245.

Bentler, P.M. (2005). *EQS 6 structural equations program manual.* Encino, CA: Multivariate software (www.mvsoft.com).

Blackorby, J., & Wagner, M. (1996). Longitudinal postschool outcomes of youth with disabilities: Findings from the National Longitudinal Transition Study. *Exceptional Children*, *62*, 399–413.

Bodin, D., Pardini, D. A., Burns, T. G., & Stevens, A. B. (2009). Higher order factor structure of the WISC-IV in a clinical neuropsychological sample. *Child Neuropsychology, 15*, 417-424.

Bollen, K. A., & Long, J. S. (1983). *Testing structural equation models*. Newbury Park, CA: Sage.

Brannick, M. T. (1995). Critical comments on applying covariance structure modeling. *Journal of Organizational Behavior, 16*, 201-13.

Brown, T. A. (2006). Confirmatory factor analysis for applied research. New York, NY: Guilford.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K.A. Bollen & J.S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Byrne, B. M. (2006). *Structural equation modeling with EQS and EQS/Windows: Basic concepts, applications, and programming* (2nd ed.). Thousand Oaks, CA: Sage.

Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications and programming*. New York, NY: Routledge.

Carroll, J. B. (1993). What abilities are measured by the WISC-III? [Special issue. Monograph, WISC-III series]. *Journal of Psychoeducational Assessment, 11*, 134-143.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*, 464-504.

Chen, Z. & Siegler, R. S. (2000). Across the great divide: Bridging the gap between understanding of toddlers' and other children's thinking. *Monographs of the Society for Research in Child Development, 65* (2), 1-96.

Chen, F. F., Sousa, K. H., & West, S. G. (2005). Testing measurement invariance of second-order factor models. *Structural Equation Modeling, 12*, 471-492.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness of fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233-255.

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management, 25*, 1–27.

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

Conger, A. J., Conger, J. C., Farrell, A. D., & Ward, D. (1979). What can the WISC-R measure? *Applied Psychological Measurement, 3*, 421-436.

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

Dean, R. (1980). Factor structure of the WISC-R with Mexican American children. *Journal of School Psychology, 18*, 234-239.

Deary, I. J., Whalley, L. J., Lemmon, H., Crawford, J. R., & Starr, J. M. (2000). The stability of individual differences in mental ability from childhood to old age: Follow-up of the 1932 Scottish Mental Survey. *Intelligence, 28*, 49-55.

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development, 43,* 121-149.

Frisby, C. L. (1999). Straight talk about cognitive assessment and diversity. *School Psychology Quarterly, 14,* 195-207.

Gignac, G. E. (2008). Higher-order models versus direct hierarchical models: *g* as superordinate or breadth factor? *Psychology Science Quarterly, 50*, 21-43.

Gorsuch, R. L. (2003). Factor analysis. In J. A. Schinka, & W. F. Velicer (Eds.), *Handbook of psychology: Research methods in psychology* (Vol. 2., pp. 143-164). Hoboken, NJ: Wiley.

Green, S. B., Thompson, M.S., & Babyak, M.A. (2010). A monte carlo investigation of methods for controlling type I errors with specification searches in strucutral equation modeling. *Multivariate Behavioral Research, 33*, 365-383.

Gregorich, S.E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups: Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care, 44,* S78-S94.

Grice, J. W., Krohn, E. J., & Logerquist, S. (1999). Cross-validation of the WISC-III factor structure in two samples of children with learning disabilities. *Journal of Psychoeducational Assessment, 17,* 236-248.

Grouzet, F. M. E., Otis, N., & Pelletier, L. G. (2006). Longitudinal cross-gender factorial invariance of the academic motivation scale. *Structural Equation Modeling, 13,* 73–98.

Gustafsson, J. E., & Undheim, J. O. (1996). Individual differences in cognitive functions. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 186-242). New York, NY: MacMillan.

Gutkin, T., & Reynolds, C. R. (1980). Factorial similarity of the WISC-R for Anglos and Chicanos referred for psychological services. *Journal of School Psychology, 18,* 34-39.

Gutkin, T., & Reynolds, C. R. (1981). Factorial similarity of the WISC-R for White and Black children from the standardized sample. *Journal of Educational Psychology, 73*, 227-231.

Hays, P. A., (2008). *Addressing cultural complexities in practice: Assessment, diagnosis, and therapy* (2nd ed., pp. 21-62). Washington, DC: American Psychological Association.

Horn, J. L. (1991). Comments on"issues in factorial invariance." In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change* (pp. 114-125). Washington, DC:American Psychological Association.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*, 117–144.

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

Hunt, E. (2011). *Human intelligence*. New York, NY: Cambridge University Press.

Individuals with Disabilities Education Improvement Act of 2004 (IDEA), Pub.L.No.108–446, 118 Stat. 2647 (2004) [Amending 20 U.S.C. § § 1400 et seq.]

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education*. Washington, DC: American Psychological Association.

Juliano, J. M., Haddad, F. A., & Carroll, J. L. (1988). Black and white, female and male children classified as learning-disabled. *Journal of School Psychology, 26*, 317-325.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141-151.

Kaufman, A. S. (1975). Factor analysis of the WISC-R at 11 age levels between 6½ and 16½ years. *Journal of Consulting and Clinical Psychology, 43*, 135-147.

Kavale, K. A., & Spaulding, L. S. (2008). Is response to intervention good policy for specific learning disability? *Learning Disabilities & Practice, 23*, 169-179.

Keith, T. A. (2005). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 581-614). New York, NY: Guilford Press.

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

Kelloway, E. K. (1995). Structural equation modeling in perspective. *Journal of Organizational Behavior, 16,* 215-24.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: Guilford Press.

Konold, T. R., Kush, J. C., & Canivez, G. L. (1997). Factor replication of the WISC-III in three independent samples of children receiving special education. *Journal of Psychoeducational Assessment, 15*, 123-137.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57*, 519 - 530.

Marsh, H. W. (1993). The multidimensional structure of academic self-concept: Invariance over gender and age. *American Educational Research Journal, 30*, 841–860.

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

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

McMahon, R. C., & Kunze, J. T. (1981). A comparison of the factor structure of the WISC and WISC-R in normal and exceptional children. *Journal of Clinical Psychology, 37,* 408 – 410.

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

Morgan, P. L., Frisco, M. L., Farkas, G., & Hibel, J. (2010). A propensity score matching analysis of the effects of special education services. *Journal of Special Education*, *43*, 236-254.

Muthén, B. & Muthén, L. (1998-2010). *Mplus User's Guide. Sixth Edition*. Los Angeles, CA: Muthén & Muthén.

National Association of School Psychologists. (2010). *Principles for professional ethics*. Retrieved from www.nasponline.org.

Osborne, R. T. (1965). Factor structure of the Wechsler Intelligence Scale for Children at preschool level and after first grade: A longitudinal study. *Psychological Reports*, *16*, 637-644.

Osborne, R. T., Anderson, H. E., & Bashaw, W. L. (1967). The stability of the WISC factor structure at three age levels. *Multivariate Behavioral Research*, *2*, 443-451.

Petersen, C. R. & Hart, D. H. (1979). Factor structure of the WISC-R for a clinic-referred population and specific subgroups. *Journal of Consulting and Clinical Psychology, 47*, 643-645.

Phelps, A., & Hanley-Maxwell, C. (1997). School to work transitions for youth with disabilities: A review of outcomes and practices. *Review of Educational Research, 67,* 197–226.

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

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25,* 111-163.

Reeve, C. L., & Bonaccio, S. (2011). On the myth and the reality of the temporal validity degradation of general mental ability test scores. *Intelligence, 39,* 255-272.

Reschly, D. (1978). WISC-R factor structures among Anglos, Blacks, Chicanos, and Native American Papagos. *Journal of Consulting and Clinical Psychology, 46*, 417-422.

Reynolds, C. R., Livingston, R. B., & Wilson, V. (2009). *Measurement and assessment in education.* Upper Saddle River, NJ: Pearson.

Reynolds, C. R., & Gutkin, T. B. (1980). Stability of the WISC-R factor structure across sex at two age levels. *Journal of Clinical Psychology, 36,* 775-777.

Roid, G. H., Prifitera, A., & Weiss, L. G. (1993). Replication of the WISC-III factor structure in an independent sample. [Special issue, Monograph, WISC-III series]. *Journal of Psychoeducational Assessment, 11,* 6-21.

Roid, G. H., & Worral, W. (1997). Replication of the Wechsler Intelligence Scale for Children – Third Edition four-factor model in the Canadian normative sample. *Psychological Assessment*, *9*, 512-515.

Rosenthal, R. & Rosnow, R. L. (1991). *Essentials of behavioral research methods and data analysis (2nd ed.).* Boston, MA: McGraw Hill.

Ryan, J. J., Glass, L. A., & Bartels, J. M. (2010). Stability of the WISC-IV in a sample of elementary and middle school children. *Applied Neuropsychology, 17,* 68-72.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test for moment structure analysis. *Psychometrika, 66*, 507–514.

Sattler, J. (1992). *Assessment of children: WISC-III and WPPSI-R supplement*. San Diego, CA: Author.

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

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. P*sychometrika, 22*, 53-61.

Schon, J., Shaftel, J., & Markham, P. (2008). Contemporary issues in the assessment of culturally and linguistically diverse learners. *Journal of Applied School Psychology, 24*, 163-189.

Silverstein, A. B. (1969). An alternative factor analytic solution for Wechsler's intelligence scales. *Educational and Psychological Measurement, 29,* 763-767.

Simonton, D. K. (2011). Exceptional talent and genius. In T. Chamorro-Premuzic, S. von Stumm, & A. Furnham (Eds.), *Wiley-Blackwell handbook of individual differences* (pp. 635-655). Malden, MA: Blackwell.

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

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25,* 331-338.

Stein, J. A., Lee, J. W., & Jones, P. S. (2006). Assessing cross-cultural differences through use of multiple-group invariance analyses. *Journal of Personality Assessment, 87,* 249-258.

Suzuki, L. A., & Valencia, R. R. (1997). Race-ethnicity and measured intelligence: Educational implications. *American Psychologist, 52,* 1103-1114.

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

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

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*, 1-10.

Tupa, D. J., Wright, M., & Fristad, M. A. (1997). Confirmatory factor analysis of the WISC-III with child psychiatric inpatients. *Psychological Assessment, 9*, 302-306.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the MI literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-69.

Wallbrown, F. H., Blaha, J., Wallbrown, J. D., & Engin, A. W. (1975). The hierarchical factor structure of the Wechsler Intelligence Scale for Children – Revised. *Journal of Psychology, 89*, 223-235.

Wang, M., Elhai, J. D., Dai, X., & Yao, S. (2010). Longitudinal invariance of posttraumatic stress disorder symptoms in adolescent earthquake survivors. *Journal of Anxiety Disorders, 26,* 263-270.

Watkins, M. W. (2006). Orthogonal higher order structure of the Wechsler Intelligence Scale for Children-Fourth Edition. *Psychological Assessment, 18*, 123-125.

Watkins, M. W. (2010). Structure of the Wechsler Intelligence Scale for Children-Fourth Edition among a national sample of referred students. *Psychological Assessment, 22*, 782-787.

Watkins, M. W., & Canivez, G. L. (2001). Longitudinal factor structure of the WISC-III among students with disabilities. *Psychology in the Schools, 38*, 291-298.

Watkins, M. W., Wilson, S. M., Kotz, K. M., Carbone, M. C., & Babula, T. (2006). Factor structure of the Wechsler Intelligence Scales-Fourth Edition among referred students. *Educational and Psychological Measurement, 66*, 975-983.

Wechsler, D. (1939). T*he measurement of adult intelligence*. Baltimore, MD: Williams & Wilkins.

Wechsler, D. (1949). *Wechsler Intelligence Scale for Children.* New York, NY: The Psychological Corporation.

Wechsler, D. (1958). The measurement of adult intelligence. Baltimore, MD: Williams & Wilkins.

Wechsler, D. (1974). *Wechsler Intelligence Scale for Children-Revised.* New York, NY: The Psychological Corporation.

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

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale-Third Edition.* San Antonio, TX: The Psychological Corporation.

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

Wechsler, D. (2002). *Wechsler Preschool and Primary Scale of Intelligence - Third 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 Intelligence Scale for Children–Fourth Edition technical and interpretive manual.* San Antonio, TX: Psychological Corporation.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.). *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281-322). Washington, DC: American Psychological Association.

Willett, J. B., Singer, J. D., & Martin, N. C. (1998). The design and analysis of longitudinal studies of development and psychopathology in context: Statistical models and methodological recommendations. *Development and Psychopathology, 10*, 395-426.

Wilson, M. S., & Reschly, D. J. (1996). Assessment in school psychology training and practice. *School Psychology Review*, *25,* 9-23.

Wright, A. J. (2011). *Conducting psychological assessment: A guide for practitioners*. Hoboken, NJ: Wiley.

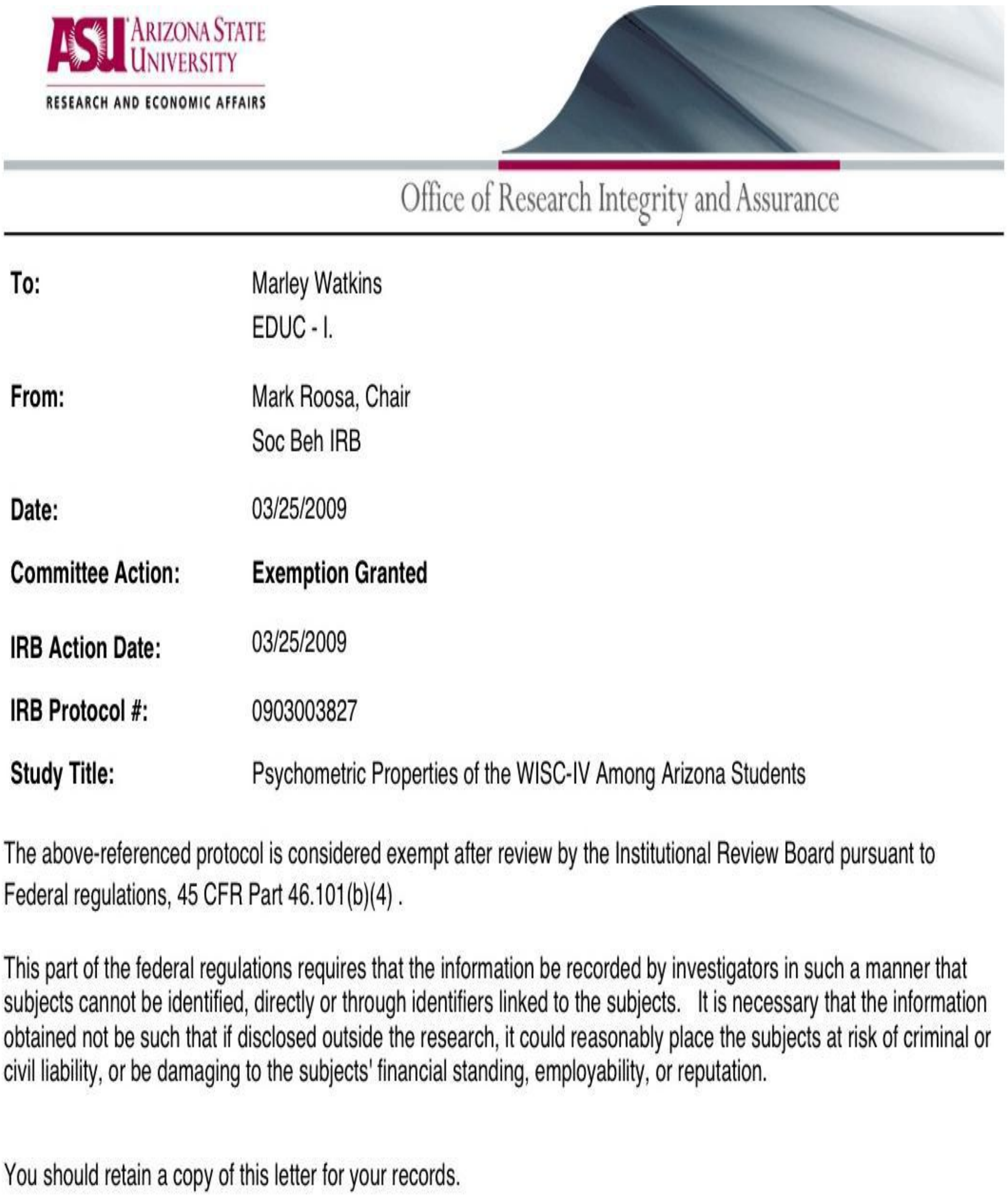
Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research & Evaluation, 12,* 1-26.

Zachary, R. A. (1990). Wechsler's intelligence scales: Theoretical and practical considerations. *Journal of Psychoeducational Assessment, 8,* 276-289.

APPENDIX

APPENDIX A

Institutional Review Board Approval of Data Collection



ASU ARIZONA STATE UNIVERSITY
RESEARCH AND ECONOMIC AFFAIRS

Office of Research Integrity and Assurance

To:	Marley Watkins EDUC - I.
From:	Mark Roosa, Chair Soc Beh IRB
Date:	03/25/2009
Committee Action:	**Exemption Granted**
IRB Action Date:	03/25/2009
IRB Protocol #:	0903003827
Study Title:	Psychometric Properties of the WISC-IV Among Arizona Students

The above-referenced protocol is considered exempt after review by the Institutional Review Board pursuant to Federal regulations, 45 CFR Part 46.101(b)(4) .

This part of the federal regulations requires that the information be recorded by investigators in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects. It is necessary that the information obtained not be such that if disclosed outside the research, it could reasonably place the subjects at risk of criminal or civil liability, or be damaging to the subjects' financial standing, employability, or reputation.

You should retain a copy of this letter for your records.