Comparison of the Factor Structure of the Reynolds Intellectual Assessment Scales (RIAS) in a Typically-Developing and Mixed Clinical Group of Canadian Children

by

Julie K. Irwin

B. A. H., University of Guelph (2007)

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Psychology

Supervisory Committee

Comparison of the Factor Structure of the Reynolds Intellectual Assessment Scales (RIAS) in a Typically-Developing and Mixed Clinical Group of Canadian Children

by

Julie K. Irwin

B. A. H., University of Guelph (2007)

Supervisory Committee

Dr. Kimberly A. Kerns, (Department of Psychology)
Supervisor

Dr. Mauricio Garcia-Barrera, (Department of Psychology)
Departmental Member

Abstract

Supervisory Committee
Dr. Kimberly A. Kerns, (Department of Psychology)
Supervisor
Dr. Mauricio Garcia-Barrera, (Department of Psychology)
Departmental Member

Objective. This thesis examines the extent to which an intelligence test, the Reynolds Intellectual Assessment Scales (RIAS), aligned with the Carroll-Horn-Cattell theory of intelligence in children ages 4-18 who are either typically-developing or who have a variety of clinical impairments. Other aspects of the RIAS's construct validity were also evaluated, including its relationship with the Wechsler Intelligence Scales for Children – Fourth Edition (WISC-IV) and whether the RIAS measures intelligence in the same way in typically-developing children as in children with traumatic brain injury (TBI).

Methods. Confirmatory factor analysis was used to evaluate the fit of one-factor (*g*) and two-factor (Verbal Ability and Non-Verbal ability) models in each sample. Configural and measurement invariance of each model were evaluated across the typically-developing group and a group of children with TBI. Correlations between scores on the RIAS and WISC-IV were examined in a group of children with clinical disorders.

Results. The two-factor model fit the data of both groups while the one-factor model provided good fit to only the typically-developing group`s data. Both models showed configural invariance across groups, measurement invariance of the two-factor model, and partial measurement invariance of the one-factor model (What`s Missing subtest unconstrained), but scalar invariance was not established for either model. RIAS's verbal subtests and indexes correlated with theoretically consistent WISC-IV indexes but the RIAS's nonverbal subtests and indexes did not correlate highly with WISC-IV performance subtests. All RIAS index scores were higher than WISC-IV index scores.

Conclusions. Evidence for the interpretability of the NIX and VIX as separate indexes was not found. The VIX is a valid index of crystallized abilities but the NIX does not adequately measure fluid intelligence. The CIX appears to provide a valid measure of *g*, but may be overly reliant on verbal abilities. The RIAS has significant validity issues that should limit its use in making important decisions.

Table of Contents

List of Tables

List of Figures

Acknowledgments

The contents of this thesis have not been published elsewhere. The results of a preliminary analysis were accepted for poster presentation at the 2011 meeting of the International Neuropsychological Society.

Some data reported in this article were collected as part of a funded project aimed at providing local normative data on the RIAS through the Queen Alexandra Centre for Children's Health (QACCH) in Victoria, British Columbia. JKI was supported in part by a Canada Graduate Scholarship Master's Award from the National Sciences and Engineering Research Council of Canada (2009-2010), by a University of Victoria President's Research Fellowship (2010-2011), and by a Petch Research Scholarship (2010-2011).

JKI wishes to thank Dr. Kimberly Kerns, Dr. Mauricio Garcia-Barrera, Dr. Michael Joschko, and Dr. Stuart MacDonald for their help in the preparation of this thesis.

Introduction

Intelligence has been a historically difficult construct to define. Experts in the field of intelligence have offered varying definitions of it, including: "to judge well, to comprehend well, to reason well" (Binet & Simon, 1916, pp. 42-43); "educing either relations or correlates" (Spearman, 1923, p.300); "the aggregate or global capacity of the individual to act purposefully, to think rationally and to deal effectively with his environment" (Wechsler, 1958, p. 7); "goal-directed adaptive behavior" (Sternberg & Salter, 1982, p. 3); and "…the degree to which, and the rate at which, people are able to learn, and retain in long-term memory, the knowledge and skills that can be learned from the environment" (Carroll, 1997, p. 44).

Attempts to operationalize the construct of intelligence have also abounded, and there exist today many tests that purport to measure intellectual abilities. Determining the validity of such tests is paramount, since they are widely used to make important decisions in clinical, vocational, educational, forensic, and social support settings. Indeed, intelligence tests are integral to the diagnosis of intellectual disabilities, giftedness, and learning disabilities, with concomitant implications for funding and resource allocation, as well as treatment recommendations accompanying these diagnoses.

The strong role of intelligence tests in these and other decisions are traditionally defended because they are able to predict varied outcomes such as: academic performance (Neisser, et al., 1996; Brody, 1992); job training performance (Hunter & Hunter, 1984); occupational level attainment and job performance (Schmidt & Hunter, 1998); ability to complete everyday tasks (Gottfredson, 1997); as well as a variety of

social and economic outcomes (Hernstein & Murray, 1994). In turn – and perhaps tautologically – the ability of intelligence tests to predict so many outcomes is touted as evidence for the tests' construct validity.

This practice of assessing the practical utility of intelligence tests has held a long tradition in academic and intellectual testing. In fact, the Binet-Simon scales (1905/1908), the first standardized intelligence tests, were developed by Alfred Binet and Théodore Simon with the explicit goal of using the scales to identify French children who would benefit from special education. These authors conceived of intelligence as a continuous variable that developed with age and learning, and would best be measured using complex tasks resembling everyday mental activities (e.g. language and reasoning; Binet, 1909/1973; Carroll, 1982). The Binet-Simon scale, with its various tasks of increasing complexity, served as the basis for the development of other individual intelligence tests (Carroll, 1982). The scale also influenced prominent psychologists like Robert Yerkes, Henry Goddard, and Lewis Terman who developed group-administered intelligence tests such as the Army Alpha and Beta Examinations (Yoakum & Yerkes, 1920) which were used to determine placement in - or discharge from - the military of over 1.75 million army recruits (Carson, 1993). Certain psychometric concepts such as "standardization" and "validation" were developed during this early period (e.g. Kelley, 1923; Thurstone, 1904/1931; Carroll, 1982), lending the tests more scientific credibility. However, although test developers often provide definitions of the construct they were attempting to measure, these early mental ability tests were not based on formal theories of intelligence. It is perhaps ironic then, that scores from intelligence tests have

themselves served as the basis on which some theories of intelligence were developed, as is the case with psychometric models of intelligence.

Psychometric approaches to the study of intelligence utilize statistical techniques to look at the patterns of variance and covariance between and among scores or items on different types of tasks. This tradition's roots lie with Sir Francis Galton who developed the statistical concepts of correlation (1886, 1889), the standard deviation, and regression to the mean. Importantly, he also began the tradition of trying to quantify intelligence and other dimensions along which individuals differ (Bulmer, 2003). In his effort to find methods of identifying the characteristics of genius, Galton collected measurements of physical size, reaction speed, and sensory acuity, taken from many people through his Anthropometric Laboratory at the International Health Exhibition in London from 1884-1890 (Bulmer, 2003). Later however, Wissler (1901), a student of James Cattell, demonstrated that individual differences in reaction speed and other simple mental tests did not correlate with college grades. Similarly, Galton's tests were examined by Binet (1905), who concluded that they lacked the complexity necessary to discriminate among individuals of differing intellectual abilities and ages. The insight that complexity is an integral component of intelligence influenced the development of many tests and theories of intelligence thereafter (e.g. Thomson, 1951; Cattell, 1963; Guttman, 1992; Jensen, 1998).

Relatively soon after the correlational method began to be used in psychology (Wissler, 1901), Robert Spearman (1904) pioneered factor analysis, one of the fundamental tools of the psychometric approach to intelligence. Interestingly, the data Spearman first used were scores on tests of discriminations of light, weight, pitch, as well

as grades in a number of academic subjects. Spearman set up a correlation matrix between academic ranks and test scores and found that the correlations could be arranged and analyzed "hierarchically." He found that the variables positively inter-correlated, but also that they appeared to measure one common factor, though each to a different degree. He published his "Two Factors of Intelligence" (1904), postulating that each test in a set measures a common, general factor (*g*), and also uniquely measures a specific factor, *s* (e.g. specific factors related to Math and English test scores), reflected by any residual variance. Spearman's work led to the widespread conceptualization of intelligence as a mostly unitary trait that reflected the presumably innate capacity of an individual to learn, to reason, and to adapt to environmental demands. This view was clearly reflected in the mental tests produced in the first 30 years after the Simon-Binet scales (1904) were developed; most intelligence tests of this era produced only a single score, usually referred to as an "intelligence quotient" (IQ; Carroll, 1982).

Today, factor analysis is generally used: i) to identify a smaller number of latent (not directly measured) factors which linearly represent the intercorrelations among a set of variables; and ii) to identify or confirm relationships (i.e. structure) between underlying factors and variables and sometimes among latent factors themselves. The calculations involved in these analyses are possible largely due to technological advances. Having developed his "factor analysis" by hand, Spearman's methods were less sophisticated by today's standards, relying on relatively simple algebraic procedures to "factor" correlation matrices; his ability to identify group factors beyond *g* was limited. It is perhaps unsurprising then, that advances in factor analytical methods themselves have broadened the scope of intelligence research beyond Spearman's *g*.

One such challenge to the view of intelligence as a unitary trait came from the work of Louis Leon Thurstone who made major contributions to factor analytic methodology. Specifically, Thurstone (1931b, 1935, 1940, 1947) developed multiple-factor analysis and introduced the concepts of: common-factor variance and communalities (common variance among tests in a set that can be analyzed into common factors); rotation to simple structure (allowing for a more psychologically meaningful description of a data set than purely mathematical descriptions had allowed for); and the concepts of correlated factors and oblique-factor structure, which would allow for further factoring when factors themselves were correlated. Thurstone (1938) applied his factor analytic methods to 56 mental test variables and interpreted seven factors (perceptual speed, verbal comprehension, spatial visualization, word fluency, number facility, reasoning, and associative memory) as being psychologically meaningful. These seven factors were postulated to represent distinct "primary mental abilities," each of which could theoretically be tested purely (something Thurstone attempted to do with his "Primary Mental Abilities Test", 1938). Thurstone argued that there was not just one type of intelligence, but that there were many kinds. Individuals could vary in their levels on each, which would be apparent on "pure" tests of that ability. Conversely, tests composed of a mélange of tasks would require the application of these underlying mental abilities, but different individuals could score similarly on such an imprecisely-designed test, even if they differed in their individual levels of underlying primary mental abilities. In fact, part of Thurstone's argument against Spearman's *g* was that it was a statistical artifact of "impure" mental tests which masked individuals' patterns of intellectual strengths and weaknesses. At the very least, it was unclear whether there was a

functional meaning of the variable that would account for the intercorrelations among tests. Nevertheless, researchers continued to find a common factor when factor analyzing groups of cognitive tests, naming the phenomenon the 'principle of positive manifold.' However, models that included multiple factors were consistently found to fit cognitive abilities data better than unitary factor models (e.g. Rimoldi, 1948; Guilford, 1967; Willoughby, 1927). That is, multiple factors, sometimes hierarchically arranged (Burt, 1949), were needed to account for the intercorrelations among cognitive abilities.

With the acknowledgement that intelligence was likely a construct of multiple factors, a desire emerged among psychologists to show that distinct patterns of covariation were, in fact, indicative of truly functionally and developmentally distinct abilities and factors. Thus, there was an increasing focus on conceiving of how different patterns of cognitive abilities and factors might be influenced by genetics, developmental factors (including neurological damage), and brain organization (Horn, 1991). This new focus led to the development of the Cattell-Horn G*f*-G*c* theory of intelligence.

The Cattell-Horn Theory of Intelligence

Raymond Cattell (1941) hypothesized that the development of cognitive abilities is influenced, firstly, by differing cultural and educational experiences, and secondly, by differences in genetic endowments and neurological development. Therefore, he thought, individual variability on cognitive ability tests was due to the influence of: variability in genetic factors (*G*); variability in the development of general ability due to environmental factors (*dG*); variability in how closely matched cultural-educational experiences are with the testing situation (*C*); iv) variability in test-specific abilities (*s*); variability in test and test-taking familiarity (*t*); variability in the application of ability due to motivational

factors (*fr*); and measurement errors (*c*). Of particular interest is Cattell's (1941) conceptualization of *G*, which was a culture-fair ability to perceive complex relations, independent of the field or subject in which it is exercised. On the other hand, *dG* and *C* were aspects of "crystallized intelligence" and were influenced by cultural and educational experiences. In turn, *dG* would be able to either impair or augment *G*, depending on the extent and quality of educational and cultural opportunities. Later, Cattell, (1943) postulated that general ability was comprised of: i) fluid ability, which is needed in new situations, for perceiving relations, and for speeded performances; and ii) crystallized ability which is apparent when relations are perceived in known material and in speeded performance.

Building on Cattell's work, John Horn (1965; Horn & Cattell, 1966) added six broad factors to the (re-named) fluid intelligence and crystallized intelligence factors. The G*f*-G*c* theory was extended even further (Horn, 1991; Horn & Stankov, 1982; Horn, Donaldson, & Engstrom, 1981; Horn & Noll, 1997) to include a total of ten factors posited. These are: Fluid Intelligence (*Gf*), the use of deliberate mental operations to solve novel problems, usually involving reasoning; Crystallized Intelligence (*Gc*), the breadth, depth, and application of acquired knowledge and skills to solving problems; Short-Term Acquisition and Retrieval (*SAR* or *Gsm*); Visual Intelligence (*Gv*); Auditory Intelligence (*Ga*); Long-Term Storage and Retrieval (*TSR* or *Glr*); Cognitive Processing Speed (*Gs*); Correct Decision Speed (CDS); Quantitative Knowledge (*Gq*); and Reading and Writing (G*rw*; Horn, 1988; McGrew, Werder, & Woodcock, 1991). Reflecting the influence of Thurstone's theory of primary mental abilities, there is no *g* factor in the extended G*f*-G*c* theory. The exclusion of a *g* factor is perhaps the most significant

difference between the G*f*-G*c* theory of intelligence and the theory it heavily influenced, John Carroll's (1993) three-stratum model of intelligence.

The Cattell-Horn-Carroll (CHC) Theory of Cognitive Abilities

In his *Human Cognitive Abilities: A Survey of Factor-Analytic Studies* (1993), John Carroll reported the results of his extensive exploratory factor analyses of over 460 cognitive ability datasets. Based on these analyses, he proposed a three-stratum theory of intelligence. This hierarchical model held that mental abilities are arranged in at least three strata, with *g* at the highest level (i.e. stratum III) and, orthogonal to *g* and to each other, several broad abilities/factors at stratum II, and a greater number of narrow abilities associated with broad factors at stratum I (Carroll, 1993; McGrew, 2005). The stratum II broad abilities of the three-stratum model are very similar to those identified in the G*f*-G*c* model. However, in contrast with the G*f*-G*c* model, Carroll: i) did not include a quantitative knowledge (G*q*) domain; ii) listed short-term memory (G*sm*) and longer-term memory and retrieval (G*lm*) under a single memory factor (G*y*); and iii) included reading and writing abilities (G*rw*) under G*c* rather than as a stratum II ability.

Though differences exist between the two theories, there is considerable overlap between Carroll's three-stratum model (1993) and Cattell-Horn's G*f*-G*c* models (1965; Horn & Blankson, 2005; Horn & Noll, 1997). Consequently, an integrated Cattell-Horn-Carroll (or CHC) model has emerged that explicitly combines both models (Daniel, 1997, 2000; McGrew, 1997, 2005, 2009; Sternberg & Kaufman, 1998; Snow, 1998). Most versions of the integrated CHC model recognize nine or ten broad abilities at stratum II (McGrew, 2009). These are: fluid reasoning or fluid intelligence (G*f*); comprehension-knowledge or crystallized intelligence (G*c*); visual processing (G*v*); auditory processing

(G*a*); short-term memory (G*sm*); long-term storage and processing (G*lr*); cognitive processing speed (G*s*); decision and reaction speed (G*t*); quantitative knowledge (G*q*); and reading and writing (G*rw*). In addition, several other broad abilities have been proposed and are under investigation, including: general (domain-specific) knowledge (G*kn*); tactile abilities (G*h*); kinesthetic abilities (G*k*); olfactory abilities (G*o*); psychomotor abilities (G*p*); and psychomotor speed (G*ps*) (McGrew, 2005).

Although other theories of intelligence exist (e.g. Sternberg, 1985; Gardner, 1993; Ceci, 1996; Guilford & Paul, 1967; Das, Naglieri, & Kirby, 1994; Campione & Brown, 1978; Borkowski, 1985), the integrated Cattell-Horn-Carroll model has been widely recognized by researchers as a useful framework with which to examine the relationships between general and specific factors, and their ability to predict outcomes (McGrew, 1997). The CHC theory has also gained popularity among cognitive test developers (Alfonso, Flanagan, & Radwan, 2005). The model has been used to assess the validity of existing tests (e.g. Wechsler series of tests), while other measures (and revisions of existing tests) have been designed explicitly to measure factors identified in the CHC model (mainly broad factors and *g*). The Reynolds Intellectual Assessment Scales (RIAS; Reynolds & Kamphaus, 2003) was one such cognitive test that was designed based on the CHC and Cattell-Horn's G*f*-G*c* models of intelligence.

The Reynolds Intellectual Assessment Scales

The Reynolds Intellectual Assessment Scales (RIAS; Reynolds & Kamphaus, 2003) was designed to provide a comprehensive measurement of intelligence for individuals aged 3-94 years. The four core RIAS subtests have a decreased reliance on reading and psychomotor speed and can be administered in 20-30 minutes, features that

have made it increasingly popular with practitioners. The test was based on the Cattell-Horn (1966) theory of intelligence but was influenced by Carroll's (1993) three-stratum, or Cattell-Horn-Carroll (CHC), model of intelligence (Reynolds & Kamphaus, 2003). The authors explicitly chose subtests that were "*g*-saturated" as these have been shown to be good predictors of various outcomes (Reynolds & Kamphaus, 2003). Furthermore, an attempt was made to select subtests that would tap fluid abilities and minimize tests of psychomotor speed as these are the best and worst measures of *g*, respectively. The RIAS' four core subtests, two verbal and two nonverbal (see Table 1), constitute three index scores: i) the Composite Intelligence Index (CIX) is composed of all four subtests and represents overall intelligence (*g*), including the ability to reason and solve problems; ii) the Verbal Intelligence Index (VIX) measures verbal reasoning and crystallized abilities; and iii) the Nonverbal Intelligence Index (NIX) assesses nonverbal reasoning and fluid reasoning abilities (Reynolds & Kamphaus, 2003).

To provide evidence of construct validity of the CIX, NIX, and VIX, the RIAS test authors performed exploratory factor analyses (EFA) and confirmatory factor analyses (CFA). These analyses have subsequently been criticized (e.g. Beaujean, McGlaughlin, & Margulies, 2009). Specifically, Reynolds and Kamphaus (2003) used EFA to first extract an unrotated factor which they interpreted as *g* (evidence for CIX). Then, in an entirely separate analysis, they performed a varimax rotation with Principal Factors analysis to extract two factors (cited as evidence for the interpretability of NIX and VIX) despite the fact that these factors were highly correlated ($r = 0.61$) and the cross-loadings of subtests on each factor were sizeable enough to make an orthogonal rotation questionable (Reynolds & Kamphaus, 2003, pp. 97 – 99; Beaujean,

McGlaughlin, & Margulies, 2009). They also used the Kaiser criterion and scree plots to determine how many factors to retain, criteria which have been criticized as being too lenient (Costello & Osborne, 2005; Frazier & Youngstrom, 2007). The test authors also used CFA to compare the relative fit of one-factor (representing CIX) and two-factor models (representing NIX and VIX). The authors did not test a model with two orthogonal factors as they had posited in their EFA analyses. Of the models they fitted, they found that an oblique two-factor model had more favourable model fit indices according to typical standards (Hu & Bentler, 1999), though they argued that these analyses provided evidence of factorial validity for the CIX as well as the NIX and VIX. In fact, it appears that the interpretability of the CIX is supported only by the author's EFA while evidence for the interpretability of the NIX and VIX is provided only by the CFA methods.

Other authors have subsequently undertaken their own factor analytic studies of the RIAS. Two groups (Dombrowski, Watkins, & Brogan, 2009; Nelson, Canivez, Lundstrom, & Hatt, 2007) used an EFA approach and, in addition to examining both orthogonal and oblique rotations with EFA, inspected higher-order factor models using the Schmid-Leiman solution (Schmid & Leiman, 1957) in samples of typically-developing individuals and referred students, respectively. In both studies, Horn's parallel analysis (Horn, 1965) and Minimum Average Partial analysis (Velicer, 1976) factor extraction criteria indicated one factor solutions. Furthermore, the results of the Schmid-Leiman procedure in both studies indicated that the higher order factor (*g*) accounted for the largest proportions of total and common variance and that, while

subtests were associated with their theoretically consistent factors, first-order factor coefficients were generally fair to poor (Dombrowski et al., 2009; Nelson et al., 2007).

EFA does not allow for the imposition of substantively meaningful constraints on the model; it is an atheoretical method. Given that the RIAS was developed based on the strongly supported CHC and G*f*-G*c* theories, using a data-driven approach (i.e. EFA) to examine the test's factor structure seems unwarranted. In contrast, CFA allows for the testing of factor models as specified by theory, and as such, has been characterized as a theory-driven approach to factor analysis. In the one study that employed CFA to study the RIAS' factor structure in three samples of referred students (kindergarten-grade 12), Beaujean and colleagues (2009) tested the relative fit of one-factor and two-factor solutions and found that the latter model provided the best fit in all three samples according to the Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI), and the Root Mean Square Error of Approximation (RMSEA) fit indices.

Whereas confirmatory factor analysis is used to assess the structural aspect of validity, factorial invariance testing is used to provide evidence for the generalizability component of validity, which is the extent to which interpretations of scores on measures of the construct generalize across groups, settings, and tasks (Messick, 1995). Reynolds & Kamphaus (2003) calculated congruence coefficients and salient variable similarity indexes to examine invariance across race and sex groupings, which has been criticized as an outdated method of invariance testing (Davenport, 1990; Beaujean, et al., 2009). While the test manual stated that the RIAS can be used in the evaluation of special populations (e.g. Learning Disabilities, ADHD, Mental Retardation, etc.), only means and standard deviations of these groups' scores were provided without an investigation of

the invariance of the factor structure between the typically-developing individuals in the normative sample and individuals with various disorders from their sample. In particular, since the RIAS is used often in forensic contexts to examine individuals with traumatic brain injury (TBI), I propose to use invariance testing to address the questions of 1) whether the same construct (intelligence) is being measured in both the typically-developing and in a group of individuals with TBI (i.e. does the RIAS have the same factorial structure for both groups?) and 2) whether the tasks comprising the RIAS operate equivalently for both groups (i.e. is the measurement model group-invariant?).

Convergent and Divergent Validity of the RIAS

RIAS index scores have been correlated with index scores from a number of established intelligence tests to provide evidence of construct validity. The RIAS Technical Manual (2003) reports moderate correlations (.71-.75) between the CIX, NIX, and VIX and the FSIQ, PIQ, and VIQ from the WAIS-III. The pattern of correlations is different when the RIAS was compared with the WISC-III (Reynolds & Kamphaus, 2003), as follows: FSIQ-CIX, .76; VIQ-VIX, .86; NIX-PIQ, .33. A similarly low (.44) correlation between the NIX and PIQ was reported by McChristian, Windsor, & Smith (2007). However, correlations between the WISC-IV and RIAS index scores were higher: CIX-FSIQ, .9; VIX-VCI, .9; NIX –PRI, .72 (Edwards & Paulin, 2007). Reynolds & Kamphaus (2009) have argued that the lower correlations of the NIX with the PIQ and PRI are due to the RIAS' decreased reliance on motor and performance speed elements.

Krach, Loe, Jones, & Farrally (2009) compared scores on the RIAS to those on the Woodcock-Johnson-III Tests of Cognitive Abilities (WJ-III-Cog; Woodcock, McGrew, & Mather, 2001) in university students. They found high correlations between

the CIX and VIX and the WJ-III-Cog measures of both *g* and crystallized intelligence (G*c*). In fact, the CIX correlated more highly with the G*c* composite than with the general intellectual ability index (*g*) in that sample. Citing the moderate correlations of NIX with the WJ-III's indexes of *g*, crystallized, and especially fluid intelligence, these authors concluded that the NIX scores are "not interpretable under the Gf-Gc or CHC frameworks" (Krach et al., 2009, p. 363) and that if information beyond crystallized and general intelligence are needed, another test should be selected.

A number of studies have found that the RIAS produces significantly higher index scores than those indicated by other intelligence tests. Higher scores on the RIAS have been found when compared with: the Woodcock-Johnson-III Test of Cognitive Ability (Krach, Loe, Jones, & Farrally, 2009) in university students; on the WISC-IV among referred students ages 6-12 years (Edwards & Paulin, 2007); and on the WAIS-III in a group of intellectually disabled individuals aged 16-57 years (Umphress, 2008). These findings are of particular concern since access to funding, community resources, and even legal decisions (e.g. eligibility for the death penalty in the United States as in *Atkins v. Virginia*) are impacted by whether individuals meet a certain threshold on intelligence tests.

Current Study and Research Questions

Prior studies addressing issues of validity of the RIAS have produced conflicting results. In order to interpret scores on the RIAS, the validity of the three index scores (CIX, NIX, and VIX) was examined; the factor structure of the RIAS and its invariance across groups (i.e. typically-developing and TBI groups) were studied. Comparisons of

the index scores with existing measures of intelligence were also made. This study aimed to address the following research questions:

1. Does a one-factor or a two-factor model fit the RIAS data better in a sample of typically-developing children and in a sample of children referred to a clinic for cognitive testing?
2. To what extent is the factor structure of the RIAS the same in a sample of typically-developing Canadian children as in a sample of Canadian children with histories of traumatic brain injury?
3. Is there evidence of convergent and divergent validity of the RIAS index scores when compared to WISC-IV index scores in a mixed clinical sample?
4. Are there significant differences between RIAS index scores and WISC-IV index scores in a mixed clinical sample?

The current study utilized CFA to compare the relative fit of one-factor (CIX) and two-factor (VIX and NIX) models to RIAS data. The data were retrospective and archival from two samples of children, ages 4-18 years. This age range is when children are typically assessed for whether or not they will receive special education and other community services, so a better understanding of the psychometric properties of the RIAS in this age range is crucial. One sample was comprised of typically-developing children, a second was of children with mixed clinical issues, and a third was comprised of children with histories of traumatic brain injuries, as described in the Methods section.

To assess the convergent and divergent validity of the RIAS, a number of comparisons were made between the RIAS index scores and several index scores from the Wechsler Intelligence Scale for Children – Fourth Edition (WISC-IV; Wechsler,

2003), which was administered to a sub-group of children in the mixed-clinical sample at approximately the same time that the RIAS was given.

Hypotheses

1. A two-factor model will fit the RIAS data better than a one-factor model in both samples, consistent with the theoretical underpinnings of the RIAS (i.e. the Cattell-Horn G*f*-G*c* model), and with findings from previous studies utilizing confirmatory factor analysis to examine the factor structure of the RIAS (Beaujean, et al., 2009; Reynolds & Kamphaus, 2003).
2. The factor structure of the RIAS will be invariant between the children with traumatic brain injury and the typically-developing sample of children.
3. The overall index scores (CIX and FSIQ), the fluid intelligence index scores (NIX and PRI), and the crystallized intelligence index scores (VIX and VCI) of the RIAS and WISC-IV will have high, positive correlations. Lower correlations should be found in comparing NIX with VCI and VIX with PRI. Similarly, RIAS verbal subtests should correlate more highly with WISC-IV subtests that comprise the VCI than with those of the PRI, while the opposite pattern should be true for nonverbal subtests of the RIAS. However, if the RIAS CIX is truly a strong measure of G*c*, all subtests will correlate more with WISC-IV VCI subtests than with PRI subtests.

Methods

Participants

Typically-developing children. Archival data for 187 typically developing children (86 female, 101 male), ages 4.08-18.83 years (*M* = 9.97 years; *SD* = 3.76) were utilized for the study. They were selected from a larger study collecting Canadian normative data for the RIAS, conducted through the Queen Alexandra Centre for Children's Health (QAACH) in Victoria, British Columbia. In addition to collecting Canadian normative data, this larger study sought to gather evidence for the construct validity of the RIAS in a typically-developing population of children, but did not originally include the use of factor analysis or planned comparisons with a clinical group of children. Participants were recruited from Vancouver Island school and community sources. Exclusionary criteria included any factors that might interfere with performance on the RIAS (e.g. colour blindness, alcohol or drug dependence, uncorrected vision or hearing loss, recent history of head injury, or current use of psychotropic medication). Participants in the larger study sample ranged in age from 3 – 22 years, but those in the lower and upper ranges, respectively, were excluded from current analyses in order to match the age range of the mixed clinical group. Participants' parents were predominately White (156 White mothers, 151 White fathers), with the remaining parents indicating the following ethnicities: Asian (25 mothers, 22 fathers); First Nations (2 mothers, 7 fathers); Black (1 father); and Hispanic (1 mother). Parental ethnicity information was missing for 3 mothers and 6 fathers (2.7%). Ethics approval was obtained from the Victoria Island Health Association Research Review and Ethical Approval Committee (Ethics approval # H2005-21) to collect these data initially and

from the Joint Victoria Island Health Association Research Review and Ethical Approval Committee and University of Victoria's Human Research Ethics Board (Ethics approval #J2011-47) for further analyses.

Mixed clinical group. Archival data for a clinical group of 164 children (68 female, 96 male), ages 4.25-18.5 (M = 12.77, SD = 3.79) were also utilized. Information about ethnicity was unavailable, though it is one QACCH neuropsychologist's impression that most patients seen at the hospital are White. Since over 85% of individuals from the Victoria Capital Region are of Caucasian ancestry (BCStats, 2011), it is likely that most of the clinical sample were White. They were referred clients at QAACH in Victoria, British Columbia who were given the RIAS as part of their neuropsychological assessment. Informed consent for using assessment data for research purposes was obtained from participants' parents or from participants who were at least 18 years old. Informed assent was obtained from participants where possible. Ethics approval for use of assessment data for research purposes was obtained from the Joint University of Victoria/ Victoria Island Health Association Research Sub-Committee Human Research Ethics Board (Ethics approval # J2011-47). Seventy-seven participants in this group were also administered the Wechsler Intelligence Scale for Children – IV (WISC-IV; Wechsler, 2003). Participants in this group had various disorders or injuries, which were grouped into six diagnostic categories for descriptive purposes. These categories are as follows:

1. Acquired brain injuries, including strokes or bleeds (n = 6), anoxic/hypoxic events (n = 2) traumatic brain injury (n = 54), and "shaken baby syndrome" (n = 1), (total n = 63)

2. Learning disabilities, including reading and math disabilities, graphomotor and visual disabilities, and nonverbal learning disabilities (n = 6)
3. Neurodevelopmental disabilities such as Attention Deficit Hyperactivity Disorder (ADHD), cerebral palsy, spina bifida, hydrocephalus, microcephaly, premature birth, seizure disorders, autism spectrum disorders, developmental delays, brain injury with viral or infectious etiology (e.g. encephalitis), and pre-natal/peri-natal exposure to insults or substances, including Fetal Alcohol Spectrum Disorders (n = 66). Note that most individuals in this group had multiple diagnoses with diverse etiologies.
4. Congenital anomalies (not described to preserve anonymity of participants with rare disorders) (n = 7)
5. Complicated neuropsychiatric referral wherein at least one psychiatric disorder (e.g. mood or anxiety disorder, Tourette's Syndrome, psychotic disorders, Adjustment Disorder, Conduct Disorder, etc.) is present in addition to at least one issue from another diagnostic category (e.g. history of brain injury, learning disability, ADHD, drug use, serious social stressors, etc.) (n = 20)
6. Uncomplicated medical disorders including an HIV-positive status, and diabetes mellitus (n = 2)

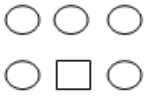
Sub-group of individuals with TBI. Data from 54 individuals (19 female), ages 6 – 18.5 years (M = 14, SD = 3.2) with traumatic brain injuries from the clinical group were used in invariance testing analyses. Information on the severity of the injuries, age of injuries, and time elapsed between injuries and testing was not

consistently available. However, injuries ranged in severity from mild to severe. There were both open and closed head injuries, number of injuries per individual ranged from one to nine (mode = 1 with only seven individuals who had sustained more than one head injury). Injuries were incurred in a diverse number of ways, including motor vehicle accidents (n = 23), sports accidents (n = 14), falls (n = 10), and assaults (n = 2) with information about how the injury was incurred unavailable for five individuals. Cognitive, behavioural, emotional, and psychological outcomes and comorbidities were also very diverse and a number of individuals appeared to have had pre-morbid conditions, including attention-deficit hyperactivity disorder and learning disabilities.

Measures

Reynolds Intellectual Assessment Scales. The Reynolds Intellectual Assessment Scales (RIAS; Reynolds & Kamphaus, 2003) is a short, individually-administered test of intelligence with normative data for use with individuals aged 3-94 years. It was designed to reduce or eliminate dependence on motor coordination, reading skills, and visual-motor speed (Reynolds & Kamphaus, 2003). The RIAS is comprised of four core subtests (see Table 1) and two supplementary memory subtests (Verbal Memory and Nonverbal Memory). The two supplementary memory subtests were not administered. The four core subtests constitute three index scores: i) the Composite Intelligence Index (CIX) is composed of all four subtests; ii) the Verbal Intelligence Index (VIX) is calculated from scores on Guess What (GWH) and Verbal Reasoning (VRZ); and iii) the Nonverbal Intelligence Index (NIX) is comprised of Odd-Item-Out (OIO) and What's Missing (WHM). Subtest scores are presented as T-scores (M = 50, SD = 10) and index scores are calculated as standard scores (M =100; SD = 15).

Table 1. ***The Reynolds Intellectual Assessment Scales (2003) subtests***

Subtest	Description	Example†	Proposed Factor
GWH	Given a set of 2-4 clues, examinees deduce the object or concept being described. Measures verbal reasoning with vocabulary, knowledge, and language development	*What is on your face and is used for smelling? Answer: nose*	Verbal
VRZ	Listen to a verbal analogy and respond with one or two words to complete the idea or proposition. Measures verbal-analytical reasoning ability.	*Sheep is to lamb as woman is to [girl].*	Verbal
OIO	Shown a card with 5-7 pictures or drawings and asked to indicate which does not belong. Measures nonverbal reasoning skills with spatial ability and visual imagery.		Nonverbal
WHM	Shown a picture with a component missing and asked to identify the missing element. Measures nonverbal reasoning (by conceptualizing the picture and analyzing its gestalt to deduce what is missing)	 [beaver is missing]	Nonverbal

† Not actual test items. *GWH* = Guess What; *VRZ* = Verbal Reasoning; *OIO* = Odd-Item-Out; *WHM* = What's Missing

The RIAS standardization sample of 2,438 people was stratified according to geographic region, educational attainment, gender, ethnicity, and age, consistent with the 2001 United States Census. The Technical Manual (Reynolds & Kamphaus, 2003) reported that internal consistency reliability coefficients ranged from .90 - .95 for the six

subtests, and from .94 - .96 for the four indexes. Inter-scorer reliability for the six subtests ranged from .95 – 1.0, while the test-retest reliability of the four index scores ranged from .83 - .91.

All participants in the typically-developing group were administered the four core subtests (GWH, VRZ, OIO, WHM) of the RIAS by a trained research assistant. Participants in the mixed clinical group were administered the RIAS by a qualified psychometrician or by a clinical neuropsychologist.

Wechsler Intelligence Scale for Children – Fourth Edition (WISC-IV). The WISC-IV (Wechsler, 2003) is an individually-administered test of children's intelligence, standardized on children ages 6:0 to 16:11 years. The test is comprised of 15 subtests, 10 of which are core subtests and 5 of which are supplemental. In the current study, only the 10 core subtests were administered to a sub-group of individuals in the clinical group. A Full Scale IQ (FSIQ) is calculated based on all scores from the 10 core subtests. In addition, subtests are combined, based on their content, to yield four index scores, as follows: Perceptual Reasoning Index (PRI) – Block Design, Matrix Reasoning, Picture Concepts; Verbal Comprehension Index (VCI) – Similarities, Comprehension, Vocabulary; Processing Speed Index (PSI) – Coding, Symbol Search; and Working Memory Index (WMI) – Digit Span, Letter-Number Sequencing. Subtest scores are converted to scaled scores (M = 10; SD = 3) while index scores are presented as standard scores (M = 100; SD = 15). Although Canadian normative data are available (Wechsler, 2004), the scores for the current study were calculated based on American norms for comparison with RIAS scores, since Canadian norms are not available for the RIAS. In the standardization sample, reliability of the Full Scale IQ score was excellent (.97),

while the reliability coefficients of the four index scores were slightly lower but still high (Perceptual Reasoning, .92; Verbal Comprehension, .94; Processing Speed, .88; and Working Memory, .92).

Statistical Analyses

All analyses were computed using IBM SPSS Statistics version 19.0.0 (SPSS Inc., an IBM company, 2010) and AMOS version 19 (Arbuckle, 2010).

Assessment of normality. Since the psychometric properties of the RIAS have not previously been examined in a Canadian sample, and to ensure that assumptions of normality were met for further statistical analyses, distributions of RIAS scaled scores and index scores were examined to assess normality. Skewness, kurtosis, Q-Q probability plots and P-P detrended probability plots, and bivariate scatterplots between index scores were examined.

Confirmatory factor analysis. Confirmatory factor analysis (CFA) was used to compare the relative fit of a one-factor model and a two-factor model to the RIAS data of the typically-developing group and the mixed clinical group, respectively. CFA allows for the testing of factor models as specified by theory. As a special case of structural equation modeling, CFA allows for the illumination of latent (unmeasured) constructs/factors (e.g. intelligence) underlying scores on various measures (e.g. IQ test) or indicators. Using this approach allows for an improvement in both the reliability and construct validity of latent constructs because CFA uses only the shared variance among indicators – attenuating for measurement error - to clarify the relationships between latent constructs. Models fitted through confirmatory factor analysis are based on *a priori* theories about the interrelations among all variables, both observed and latent. However,

unlike with structural equation modeling, no directional (causative) pathways are asserted between latent constructs, though they are allowed to co-vary.

Proposed models. The one-factor model posits that a single unitary construct, *g*, underlies scores on all four subtests of the RIAS. If this model has good fit to the data, it would provide evidence for the interpretability of the CIX. The two-factor model is comprised of a nonverbal/fluid intelligence factor, which underlies performance on the nonverbal OIO and WHM subtests, and a verbal/crystallized intelligence factor, which underlies scores on the verbal VRZ and GWH subtests. Interpretation of the NIX and VIX scores would be supported if the two-factor model fits the data well. Note that a higher-order factor model representing strata two and three of the CHC model cannot be fitted because there are too few indicators. Such a model would be underidentified without constraining parameter estimates that are not theoretically warranted (Brown, 2006).

Model estimation. All analyses were completed using the AMOS v.19.0.0 (Arbuckle, 2010) software module and IBM SPSS Statistics v.19.0.0 (SPSS Inc., an IBM Company, 2010). Maximum likelihood was used to estimate unknown parameters from sample variance-covariance matrices**.** Instead of raw scores, *T*-scores were analyzed so that scores were scaled to the same metric across the samples' age ranges. To scale model variances, a single indicator was fixed to 1.0 for each factor. Various fit criteria (see Table 2 for cut-off scores) were used to evaluate model fit and degree of parsimony, including the comparative fit index (CFI; Bentler, 1990), the chi-square goodness-of-fit test (Loehlin, 1998), the ratio of chi-square to degrees of freedom (Bollen, 1989), and the root mean square error of approximation (RMSEA; Steiger, 1990). In addition, the chi-

square difference between the two models was calculated to determine if their fits were significantly different from each other.

Table 2. ***Description of fit criteria for confirmatory factor analyses***

Fit criteria	Description	Cut-off score/range
Chi-Square Test	Evaluates the null hypothesis that an overidentified (reduced) model fits the data as well as a saturated model	A higher *p*-value indicates better fit between the specified model and the observed data.
Comparative Fit Index (CFI)	Compares relative fit of model-implied matrix to independence model matrix	Ranges from 0 to 1 with values of .95 or higher indicating good fit.
Root Mean Square Error of Approximation (RMSEA)	Estimates lack of fit per estimated parameter compared to saturated model, taking model complexity (degrees of freedom) into account	Good fit: ≤ .05 Adequate fit: .05 - .08 Borderline fit: .08 - .10 Poor fit: >.10
Chi-Square/*df*	This ratio is the minimum sample discrepancy divided by degrees of freedom. As a parsimony index, it shows how much the fit of a model to the data has been reduced by dropping one or more parameter estimates	Ratio as low as possible using fewer degrees of freedom (useful when comparing two or more specified models)

Invariance testing. In the context of multigroup CFA, factorial invariance testing allows for the examination of configural invariance and measurement invariance. Establishing configural and measurement invariance indicates that the test measures the same construct in the same way across groups; factorial invariance is necessary if cross-

group comparisons are to be made (Dimitrov, 2010). Before invariance testing can be completed, a baseline model for comparison must be estimated for each group separately. The baseline model selected for each group is the one that is the most meaningful, parsimonious, and has the best fit to the group's data (Jöreskog, 1971; Byrne, 2001). Configural invariance is demonstrated when the pattern of free and fixed model parameters is the same for both groups (e.g. the same indicators define the same latent factors). Measurement invariance includes metric invariance (factor loadings are equal across groups), scalar invariance (item intercepts are equal across groups), and invariance of item uniquenesses (across groups, item error variances/covariances are equal; Dimitrov, 2010). Invariance of these latter error parameters has been recognized as overly restrictive (Bentler, 2004) except in the case where equivalent reliability across groups is being tested (Byrne, 2004). To demonstrate weak measurement invariance, metric invariance must be established. In this case, equal factor loadings across groups allow for the comparison between latent factors and external variables since a one-unit change would be equivalent in each group. However, with weak invariance, factor means could not be compared between groups since the origin of the scale may differ for each group. Establishing strong measurement invariance requires both metric and scalar invariance; equal factor loadings and equal indicator intercepts, or means, must be shown across groups. With strong measurement invariance, factor means may be compared across groups. Item bias may be indicated by a lack of invariant intercepts. Finally, with strict measurement invariance, metric and scalar invariance, and invariance of item uniquenesses must all be evident. This type of measurement invariance indicates that

items were measured in the same way in each group; group differences on items/scales are due solely to group differences on common factors.

The forward approach (or sequential constraint imposition; Jöreskog, 1971; Byrne et al., 1989; Dimitrov, 2010) to factorial invariance testing will be employed. Moving from the least to the most constrained model, a series of nested constrained models (invariance assumed) and unconstrained models (no invariance assumed) will be compared using the chi-square difference test as more parameters (e.g. factor loadings, etc.) are constrained to be equal. Since the chi-square difference test may be overly sensitive to sample size, a difference in the Comparative Fit Index (CFI) between any two nested models will also be examined. A difference in the CFI of less than 0.01 and a non-significant chi-square value will indicate invariance of whichever parameter(s) has been constrained to be equal in the more constrained model (Cheung & Rensvold, 2002).

RIAS and WISC-IV comparisons. Correlations between pairs of conceptually similar and dissimilar index scores on the RIAS and WISC-IV were calculated to provide evidence of convergent and divergent validity, respectively, of the RIAS index scores in the mixed-clinical group. Paired-difference *t*-tests between each pair were also calculated to determine whether there were significant differences in index scores on the RIAS versus the WISC-IV. The pairs of index scores compared are as follows: CIX-FSIQ; VIX-VCI; VIX-PRI; NIX-PRI; and NIX-VCI.

Results

Data Cleaning

Missing data. One participant from the clinical group (age 16.5 years, male, diagnostic category 5) was only administered the NIX subtests so his data were excluded from analyses. There were no missing data in the typically-developing group. Index scores of ≤ 40 were replaced with values of 40 which allowed for statistical analyses but resulted in a restricted range.

Univariate Outliers. Frequency tables and histograms were examined to identify possible univariate outliers in each group's RIAS *T*-scores and index scores.

In the typically-developing group, there was a large degree of variability in scores with ranges of 48, 52, 53, and 62 for the GWH, OIO, VRZ, and WHM *T*-scores, respectively, and of 67, 77, and 79 for the VIX, NIX, and CIX index scores, respectively. Even though the distributions were wide, no *T*-scores were relatively extreme since all scores were within five points of each other. Similarly, all index scores were within six points of each other except for the highest NIX score (159) which was almost a standard deviation (14 points) higher than the next highest NIX score and the second highest CIX (150) which was 11 points higher than the third highest CIX (139). These higher scores were obtained by a male aged five years, 11 months, and a female, aged four years, one month. Examination of index scores of participants aged four - five revealed that 41/69 (59.4%) index scores in this age range were at least one standard deviation above the mean. That is, the index scores in this age range were unexpectedly high, which may be reflective of the finding that Canadian children tend to score higher than American children on standardized intelligence tests (e.g. Wechsler, 2004). The data from the two

participants with the highest index scores were not excluded since it is likely that they represent the high end of a sample with generally higher scores.

In the clinical group, there were a substantial number of participants with very low *T*-scores and index scores. There were also a number of relatively higher scores, reflected in the ranges of 66, 57, 59, and 71 for the GWH, OIO, VRZ, and WHM *T*-scores, respectively, and of 98, 89, and 94 for the VIX, NIX, and CIX index scores, respectively. However, since there were a number of these low scores, they were not technically "outliers." Even the lowest possible index score of ≤ 40 was obtained by 3.1% of the sample. At either end of the distribution of any index score or *T*-score in the clinical group, there was never a difference greater than 10 between each score except for an eleven point difference between the second highest and third highest CIX scores. That is, the clinical sample had a large range in scores and had a greater number of extremely low scores and a few relatively higher scores, but no scores were extreme relative to the sample's scores. Given the large range of scores in the typically-developing sample, it is not unexpected that there was an even larger range in a heterogeneous clinical group. Nonetheless, examination of the diagnoses of individuals with very low or relatively high scores were examined. Lowest-scoring participants (any index scores between ≤ 40 and 50) were referred with global developmental delay, Down's Syndrome, cerebral infarctions, autistic disorder, anxiety disorders, and seizure disorders while very high-scoring participants (CIX ≥ 120) were characterized by head injuries and/or ADHD with or without *in utero* drug exposure. Since these disorders fall within the purview of the RIAS, as defined in the manual (Reynolds & Kamphaus, 2003), and since they seem to be representative of some portion of a general clinical population, all data were retained.

Multivariate Outliers. When CFA analyses were completed with the total mixed clinical sample, Mardia's coefficient was 4.059 (critical ratio = 3.74). Values less than 10 are desirable and indicate multivariate normality. However, one case (age = 7.5 years, male, diagnostic category 5) had a Mahalanobis d^2 value of 21.553 ($p1$=.000; $p2$=.039) with 14.945 as the next largest value. Further examination revealed that this participant had a CIX score of 71 but a standard score difference of 60 between his/her NIX (104) and VIX (44), a 4 standard deviation difference, the largest split in the entire dataset. Furthermore, within the NIX, the OIO *T*-score was 38 and the WHM *T*-score was 65. The client had been referred for assessment with a trauma spectrum disorder and "extreme shyness and anxiety." It is unknown whether this anxiety negatively impacted the child's performance on the RIAS, especially when verbal responses were required. To assess the impact of this case on overall analyses, this outlier was removed and the CFA analyses were re-done. Mardia's coefficient became 2.651 (critical ratio = 2.435), the chi-square value of the one-factor model became non-significant, all other fit indices improved for the one-factor model (e.g. RMSEA fell from .120 to .101), and more variance in WHM *T*-scores was accounted for in both models (squared multiple correlations changed from .41 to .47 and .50 to .55 for the one- and two-factor models, respectively). Since this case had a strong impact on CFA analyses and because it is unknown whether the RIAS score can be validly interpreted, these data were excluded from further analyses.

In the typically-developing group, Mardia's coefficient of multivariate kurtosis was 2.36, critical ratio = 2.329. Maximum likelihood estimation was utilized in confirmatory factor analyses since the data were normally distributed.

Assessment of normality, linearity, and homoscedasticity. Univariate normality was assessed by examining skewness, kurtosis, frequency tables, histograms, Q-Q probability plots, and detrended probability plots for each RIAS *T*-score and index score in both groups (see Tables 3 and 4 for descriptive statistics). Linearity and heteroscedasticity were assessed by visually inspecting bivariate plots between each pair of *T*-scores and ensuring that scores clustered in a roughly oval and/or linear shape that lacked obvious bulges.

Typically-developing group. All variables in the typically-developing group were normally distributed. Skewness values of *T*-scores for each subtest ranged from -.181 – .370 (*S.E.* = .178) and kurtosis values ranged from -.057-.629 (*S.E.* = .354), all *p*-values >.05. Skewness values of index scores ranged from .242 – .384 (*S.E.* = .178) and kurtosis values ranged from .003-.537 (*S.E.* = .354), all *p*-values >.05. Q-Q probability plots and detrended probability plots indicated that the data were approximately normally distributed for all scores, though the highest GWH, OIO, and VRZ *T*-scores and NIX and CIX index scores deviated to some extent from normality. However, often there were only one or two values that deviated from normality and these were retained for the reasons described above. Inspection of bivariate plots between each pair of *T*-scores revealed approximate linearity and homoscedasticity.

Table 3

Descriptive statistics of clinical sample's RIAS scores

	n	*M* (*S.E.*)	S.D.	Skewness (*S.E.*)	Kurtosis (*S.E.*)	1	2	3	4
1. GWH *T*sc	162	43.9 (.91)	11.6	-.765 (.191)*	1.780 (.379)*	1	.788**	.601**	.616**
2. VRZ *T*sc	162	41.9 (.89)	11.4	-.427 (.191)*	.555 (.379)		1	.585**	.562**
3. OIO *T*sc	162	46.8 (.87)	11.1	-.843 (.191)*	.313 (.379)			1	.547**
4. WHM *T*sc	162	44.7 (1.09)	13.9	-.616 (.191)*	.317 (.379)				1
NIX	162	94.2 (1.47)	18.7	-.896 (.191)*	.583 (.379)				
VIX	162	90.9 (1.32)	16.8	-.719 (.191)*	1.794 (.379)*				
CIX	162	91.9 (1.42)	18.1	-.890 (.191)*	1.148(.379)*				

*$p < .05$; **$p < .01$ Note: Values rounded to the nearest tenth. *T*sc = *T*-score; GWH = Guess What; VRZ = Verbal Reasoning; OIO = Odd-Item-Out; WHM = What's Missing; NIX = Nonverbal Intelligence Index; VIX = Verbal Intelligence Index; CIX = Composite Intelligence Index

Clinical group – full sample. After removal of one multivariate outlier (described above), significant skewness was found for all RIAS *T*-scores and index scores, *z*-scores between -4.69 and -2.24, *p*-values < .05. Distributions with significant kurtosis were the GWH *T*-score, $z = 4.7$, $p < .05$, VIX, $z = 4.73$, $p < .05$, and CIX, $z = 3.03$, $p < .05$. Examination of Q-Q probability plots and detrended probability plots revealed that values in the lower ranges tended to deviate from normality for all index and *T*-scores except the VRZ *T*-score. However, Tabachnick and Fidell (2007) note that "in a large sample, a variable with statistically significant skewness often does not deviate enough from normality to make a substantive difference in the analysis" (Tabachnick & Fidell, 2007, p. 80). They contend that the actual size of the skewness and the visual appearance of the distribution are more important than the significance level. Data for this group were not transformed because: 1) No *T*-scores or index scores had skewness values greater than 1 and visual inspection of distributions revealed approximately normal distributions (Tabachnick & Fidell, 2007); 2) Statistically significant kurtosis values were all positive and underestimates of variance associated with positive kurtosis disappear with sample sizes of at least 100 (Waternaux, 1976); 3) Subtest *T*-scores and index scores are in meaningful metrics which would be difficult to interpret if data were transformed. Mardia's coefficient of multivariate kurtosis was 2.36, critical ratio = 2.329, indicating multivariate normality. Inspection of bivariate plots between each pair of *T*-scores revealed no obvious issues with linearity or homoscedasticity.

Table 4

Descriptive statistics of typically-developing sample's RIAS scores

	n	*M* (*S.E*)	S.D.	Skewness (*S.E*)	Kurtosis (*S.E*)	1	2	3	4
1. GWH *T*sc	187	56.2 (.60)	8.2	-.181(.178)	.451(.354)	1	.572**	.291**	.212**
2. VRZ *T*sc	187	55.3 (.75)	10.2	-.106(.178)	-.057(.354)		1	.292**	.167*
3. OIO *T*sc	187	57.5 (.60)	8.2	.370(.178)	.629(.354)			1	.161*
4. WHM *T*sc	187	54.1 (.82)	11.2	-.109(.178)	.292(.354)				1
NIX	187	111.5 (.99)	13.5	.384(.178)	.537(.354)				
VIX	187	110.9 (.99)	13.5	.242(.178)	.003(.354)				
CIX	187	111.9 (.93)	12.7	.315(.178)	.444(.354)				

*$p < .05$; ** $p < .01$ Note: Values rounded to the nearest tenth. *T*sc = *T*-score; GWH = Guess What; VRZ = Verbal Reasoning; OIO = Odd-Item-Out; WHM = What's Missing; NIX = Nonverbal Intelligence Index; VIX = Verbal Intelligence Index; CIX = Composite Intelligence Index

Descriptive Statistics

See Tables 3 and 4 for means, standard deviations, and zero-order correlations of the clinical and typically-developing samples' RIAS scores. One-way ANOVAs were performed to examine whether the two groups' mean *T*-scores and index scores were significantly different from each other. In all pair-wise comparisons, the typically-developing group scored significantly higher than the clinical group, as follows: GWH *T*-score, $\Delta M = 12.33$, $F(1, 347) = 133.13$, $p < .001$; OIO *T*-score, $\Delta M = 10.67$, $F(1, 347) = 105.41$, $p < .001$; VRZ *T*-score, $\Delta M = 13.28$, $F(1, 347) = 132.83$, $p < .001$; WHM *T*-score, $\Delta M = 9.36$, $F(1, 347) = 48.54$, $p < .001$; VIX, $\Delta M = 19.89$, $F(1, 347) = 150.46$, $p < .001$; NIX, $\Delta M = 17.34$, $F(1, 347) = 100.35$, $p < .001$; CIX, $\Delta M = 20.8$, $F(1, 347) = 158.272$, $p < .001$. Using a 95% confidence interval, none of the *T*-score or index score ranges overlapped between groups.

Confirmatory Factor Analyses

Model fit - typically-developing sample. The one-factor model fit the data well (see Figure 1a), $\chi^2(2, N=187) = 1.237$, $p = .539$, CFI = 1.0, RMSEA = 0 (90% C.I. = 0 – 0.126), χ^2/df ratio = .619. The two-factor model also fit the data well (see Figure 1b), χ^2 $(1, N = 187) = .380$, $p = .538$, CFI = 1.0, RMSEA = 0 (90% C.I. = 0 – 0.164), χ^2/df ratio = .380. Comparison of fit indices between models indicated that the models fit the data equally well, though the two-factor model was slightly more parsimonious than the one-factor model, according to the χ^2/df ratio values of each. Finally, the chi-square difference test indicated that the two models were not significantly different from each other, $\chi^2_D(1) = .857$, *ns*.

Model estimates - typically-developing sample.

One-factor model estimates. See Figure 1a for standardized regression weights and squared multiple correlations representing the proportion of variance accounted for in each indicator by the corresponding factor. All of the indicators (i.e. four subtests) loaded significantly onto the single factor (*g*), with critical ratios between 2.97 – 4.99, all *p*-values < .05. Standardized regression weights are as follows: GWH = .777; VRZ = .732; OIO = .391; WHM = .264. According to some conventions (e.g. Hair, et al., 1998), GWH and VRZ had "high" loadings (i.e. > 0.6) on *g*, while OIO and WHM had "low" loadings on *g* (i.e. < 0.4).

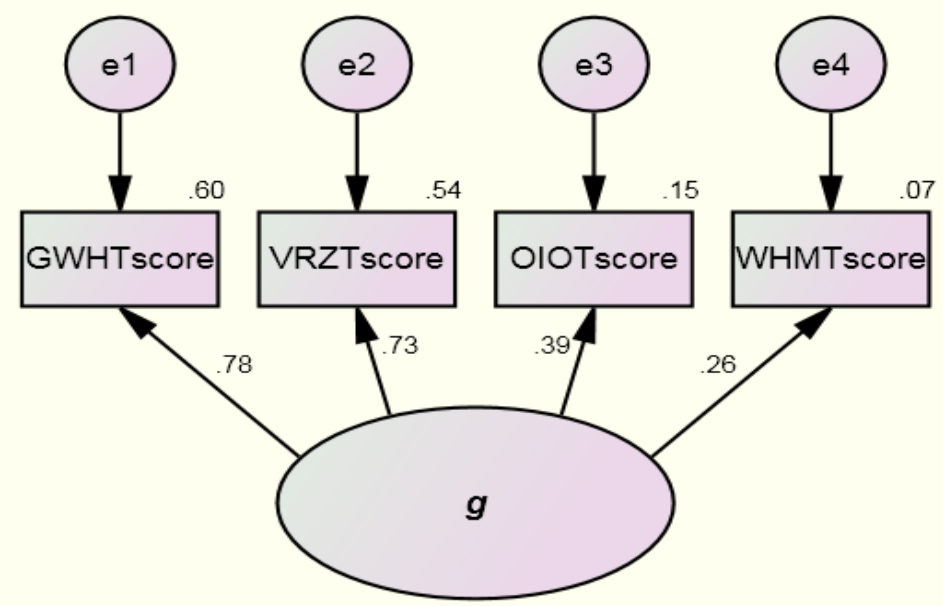


Figure 1a. One-factor model of the RIAS fit to the typically-developing group's data
Note: Standardized regression estimates are reported on straight arrows, factor covariances are on the curved arrow, and squared multiple correlations are on the upper right hand corner of indicator boxes. GWH = *Guess What*; VRZ = *Verbal Reasoning*; OIO = *Odd-Item-Out*; WHM = *What's Missing*; e = *error variances*

Two-factor model estimates. See Figure 1b for standardized regression weights and squared multiple correlations. The verbal and non-verbal factors correlated

highly, $r = 0.78$. The verbal indicators (GWH and VRZ) and nonverbal indicators (OIO and WHM) loaded significantly onto the corresponding verbal and nonverbal factors, respectively, with critical ratios ranging from 2.67 – 4.783, all *p*-values < .05. The standardized regression weights were as follows: GWH = .78; VRZ = .733; OIO = .493; WHM = .325. Both GWH and VRZ loaded highly on the verbal factor, while OIO had a moderate loading on the nonverbal factor, and WHM loaded "low" on the nonverbal factor.

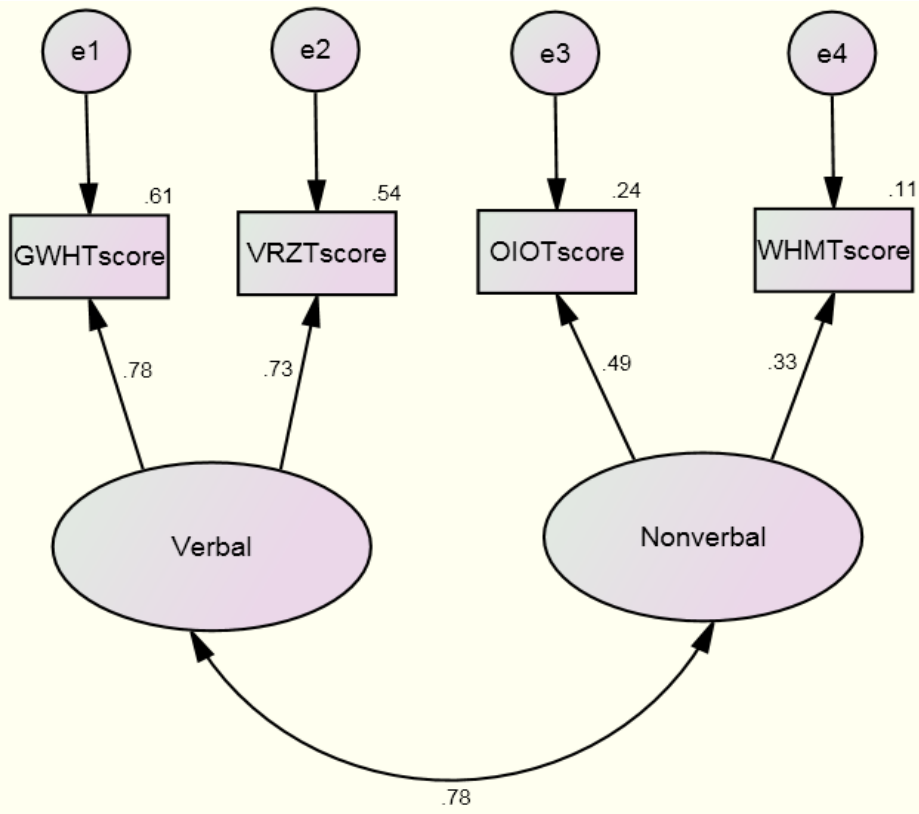


Figure 1b. Two-factor model of the RIAS fit to the typically-developing group's data. *Note:* Standardized regression estimates are reported on straight arrows, factor covariances are on the curved arrow, and squared multiple correlations are on the upper right hand corner of indicator boxes. GWH = *Guess What*; VRZ = *Verbal Reasoning*; OIO = *Odd-Item-Out*; WHM = *What's Missing*; e = *error variances*

Model fit – mixed clinical sample. According to some fit indices, the one-factor model fit the data well (see Figure 2a), $\chi^2(2, N = 162) = 5.267$, $p = .072$, CFI = .990, χ^2/df ratio = 2.633. However, the RMSEA indicated a "poor" fit of the model to the data, RMSEA = .101 (90% C.I. = 0 – 0.210), though notably, this interval included values indicating a very good fit and was close to having a “reasonable” fit by some authors’ suggestions (Sugawara & MacCallum, 1993; Jöreskog & Sörbom, 1993). As well, the PCLOSE index, which evaluates the null hypothesis that the RMSEA is .05 (i.e. a close-fitting model), was .152, indicating that the fit of the model was close.

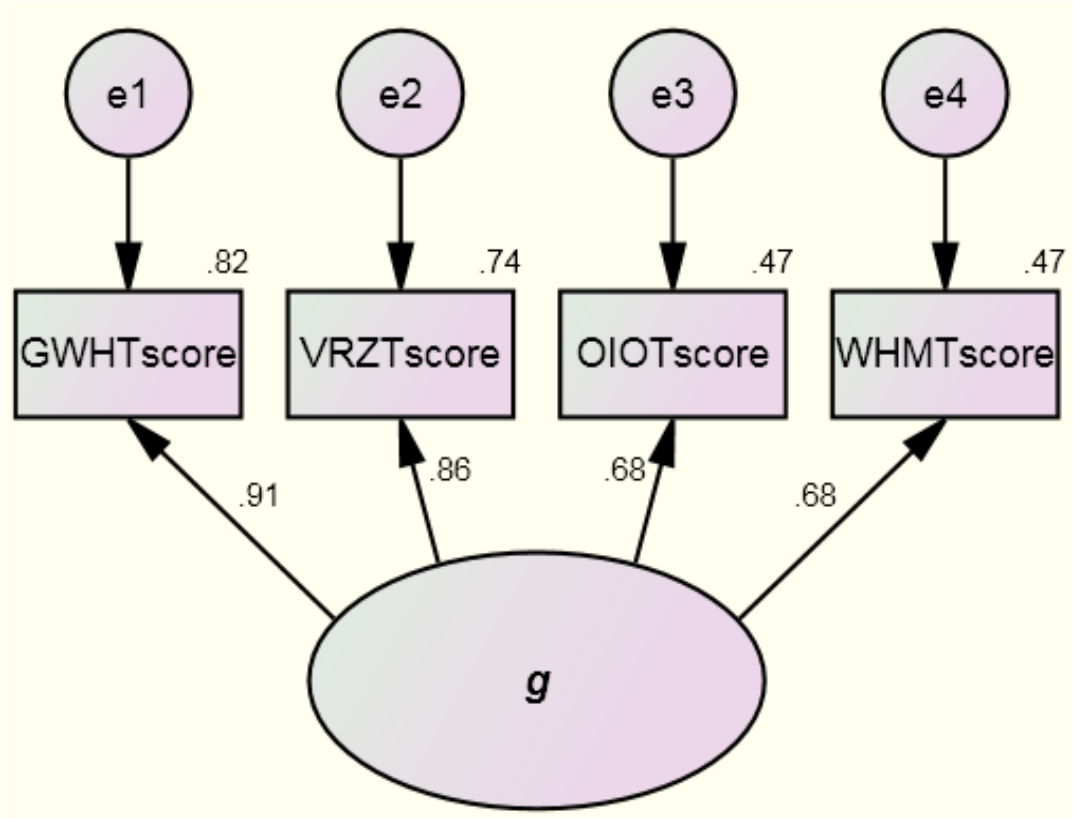


Figure 2a. One-factor model of the RIAS fit to the mixed clinical group's data

Note: Standardized regression estimates are reported on straight arrows, factor covariances are on the curved arrow, and squared multiple correlations are on the upper right hand corner of indicator boxes. GWH = *Guess What*; VRZ = *Verbal Reasoning*; OIO = *Odd-Item-Out*; WHM = *What’s Missing*; e = *error variances*

The two-factor model fit the data well according to all fit indices (see Figure 2b), χ^2 (1, N = 162) = .582, p = .538, CFI = 1.0, RMSEA = 0 (90% C.I. = 0 – 0.189), χ^2/df ratio = .582. Comparison of fit indices between models indicated that the two-factor model fits the data better than the one-factor model in the clinical group. The chi-square difference test also indicated that the two models were significantly different from each other, χ^2_D (1) = 4.685, $p < .05$. In addition, the two-factor model was more parsimonious than the one-factor model, according to the χ^2/df ratio.

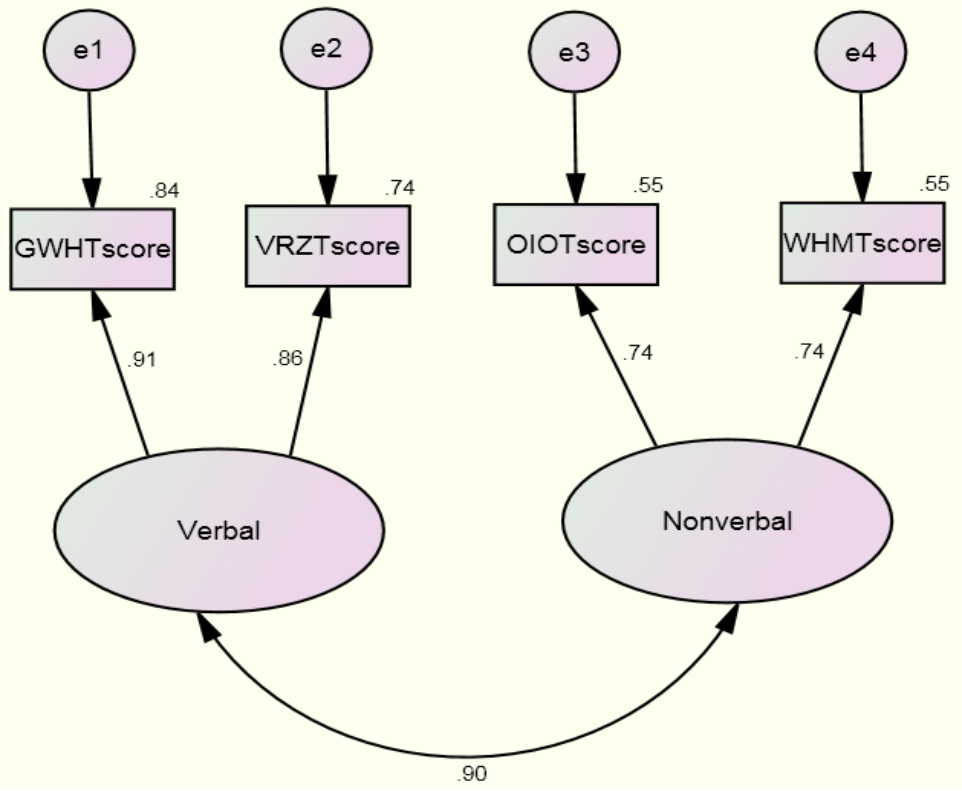


Figure 2b. Two-factor model of the RIAS fit to the mixed clinical group's data.

Note: Standardized regression estimates are reported on straight arrows, factor covariances are on the curved arrow, and squared multiple correlations are on the upper right hand corner of indicator boxes. GWH = *Guess What*; VRZ = *Verbal Reasoning*; OIO = *Odd-Item-Out*; WHM = *What's Missing*; e = *error variances*

Model estimates - mixed clinical sample.

One-factor model estimates. See Figure 2a for standardized regression weights and squared multiple correlations, representing the proportion of variance accounted for in each indicator by the corresponding factor. All of the indicators loaded significantly onto the single factor (*g*), with critical ratios between 9.822 – 13.485, all *p*-values < .05. All subtests had high loadings on *g*. Standardized regression weights are as follows: GWH = .906; VRZ = .862; OIO = .683; WHM = .683.

Two-factor model estimates. See Figure 2b for standardized regression weights and squared multiple correlations. The verbal and non-verbal factors correlated highly, $r = 0.90$. The verbal indicators (GWH and VRZ) and nonverbal indicators (OIO and WHM) loaded significantly onto the corresponding verbal and nonverbal factors, respectively, with critical ratios ranging from 8.363 – 13.044, all *p*-values < .05. The standardized regression weights were as follows: GWH = .915; VRZ = .862; OIO = .739; WHM = .740. GWH and VRZ loaded highly on the verbal factor, while OIO and WHM had high loadings on the nonverbal factor.

Invariance Testing

Descriptive statistics and normality of TBI sample. See Table 5 for descriptive statistics of TBI sample.

Table 5

Descriptive statistics of traumatic brain injured sub-sample's RIAS scores

	n	*M* (*S.E.*)	S.D.	Skewness (*S.E.*)	Kurtosis (*S.E.*)	1	2	3	4
1. GWH *T*sc	54	45.8 (1.11)	8.1	.460 (.325)	2.512 (.639)*	1	.731**	.277*	.629**
2. VRZ *T*sc	54	44.09 (1.11)	8.2	-.466 (.325)	-.320 (.639)		1	.333*	.643**
3. OIO *T*sc	54	49.44 (1.37)	10.1	-1.304 (.325)*	2.301 (.639)*			1	.361**
4. WHM *T*sc	54	49.13 (1.49)	10.9	.079 (.325)	.256 (.639)				1
NIX	54	100.1 (2.06)	15.1	-.967 (.325)*	1.914 (.639)*				
VIX	54	93.85 (1.56)	11.4	.102 (.325)	1.577 (.639)*				
CIX	54	95.98 (1.83)	13.5	-.611 (.325)	2.810(.639)*				

*$p < .05$; **$p < .01$ Note: Values rounded to the nearest tenth. *T*sc = *T*-score; GWH = Guess What; VRZ = Verbal Reasoning; OIO = Odd-Item-Out; WHM = What's Missing; NIX = Nonverbal Intelligence Index; VIX = Verbal Intelligence Index; CIX = Composite Intelligence Index

The OIO *T*-score was significantly negatively skewed, $z = -4.01$, $p < .05$, but all other *T*-scores were not skewed. The OIO *T*-score had positive kurtosis, $z = 3.6$, $p < .05$ and so did the GWH *T*-score, $z = 3.93$, $p < .05$. None of the data were transformed since the scores were in a meaningful metric and absolute skewness and kurtosis values were small.

See Table 6 for nested model descriptions and difference statistics. The factor invariance of the one-factor model between the traumatic brain injury and typically-developing groups was first evaluated despite the equivocal results when its fit was estimated separately in the whole clinical sample. In a multigroup context, the one-factor model (Model A1) fit the data of the TBI and typically-developing groups well, $\chi^2(4) = 2.458$, $p = .652$, $\chi^2/df = .614$, CFI = 1, RMSEA = 0 (90% C.I. = 0 - .078), indicating configural invariance between the groups. However, when factor loadings were constrained to be equal (Model B1), the model fit adequately, $\chi^2(7) = 13.53$, $p = .06$, $\chi^2/df = 1.933$, CFI = .963, RMSEA = 0.62 (90% C.I. = 0 – .112) but was significantly different than the unconstrained model, $\Delta\chi^2 = 11.072$, $\Delta df = 3$, $p = .011$, ΔCFI = .037, indicating a lack of full metric invariance. Modification indices indicated that the WHM factor loading was a likely source of misfit. Indeed, the WHM factor loading in the typically-developing group was .264 but was .752 in the TBI group. The factor loadings of GWH, VRZ, and OIO in the typically-developing group were .777, .732, and .391, respectively and were .837, .867, and .384, respectively in the TBI group. Thus, the WHM factor loading had the largest difference between groups. After allowing this parameter to estimate freely, the model fit well, $\chi^2(6) = 2.864$, $p = .862$, $\chi^2/df = .477$, CFI = 1, RMSEA

= 0 (90% C.I. = 0 - .050) and was not significantly different from Model A1, $\Delta\chi^2 = 1.132$, $\Delta df = 1$, $p = .287$, $\Delta CFI = 0$, establishing partial metric invariance (Model C1).

Table 6

Invariance testing steps and results across the TBI and typically-developing groups

Baseline Model	Constrained Parameters	Compared	Type of invariance Model	$\chi^2(df)$	$\Delta\chi^2(\Delta df)$	RMSEA (90% C.I.)	CFI	Δ CFI
			One-Factor Model					
Model A1	None (Unconstrained)	-	Configural	2.46(4)	-	0 (0 - .078)	1.0	-
Model B1	Factor loadings	A1	Metric	13.53(7)	11.072(3)*	.062(0-.112)	.963	.037
Model C1	Factor loadings of OIO, VRZ, and GWH	A1	Partial Metric	2.864(6)	.406(2)	0(0-.05)	1	0
Model D1	Factor loadings and item Intercepts	C1	Scalar	69.56(10)*	66.69(4)*	.158(.12-.19)	.663	.337
			Two-Factor Model					
Model A2	Unconstrained	-	Configural	.64(2)	-	0(0-.091)	1	-
Model B2	Factor loadings	A2	Metric	4.249(4)	3.609(2)	.016(0-1)	.999	.001
Model C2	Factor loadings and item intercepts	B2	Scalar	66.778(8)*	62.528(4)*	.175 (.138-.215)	.668	.331

*$p < .001$
Note: Invariance is indicated by a nonsignificant $\Delta\chi^2$ test and/or a ΔCFI of -.01 or less

To test for scalar invariance, the measurement intercepts were constrained to be equal (Model D1) and this model was compared to Model C1. Model D1 fit poorly, $\chi^2(10) = 69.556$, $p < .001$, $\chi^2/df = 6.956$, CFI = .663, RMSEA = .158 (90% C.I. = .124 - .194). There was a significant difference between Model C1 and D1, $\Delta\chi^2 = 66.691$, $\Delta df = 4$, $p < .001$, ΔCFI = .337. Modification indices did not indicate that any parameters could be changed to improve the fit. Comparison of intercepts between groups revealed that all seemed discrepant, with the typically-developing group's intercepts systematically higher than those of the TBI group. Since partial scalar invariance could not be established, investigation of more constrained models was not done.

The factorial invariance of the two-factor model between the clinical and typically-developing groups was also evaluated since this model had good fit indices when estimated separately in each larger sample. Consistent with the results of the confirmatory factor analyses, the unconstrained model (Model A2) fit the data well in the multigroup context, $\chi^2(2) = .640$, $p = .726$, $\chi^2/df = .32$, CFI = 1.0, RMSEA = 0 (90% C.I. = 0 - .091), providing evidence of configural invariance. The measurement model (Model B2; all factor loadings constrained to be equal) also fit the data well, $\chi^2(4) = 4.249$, $p = .373$, $\chi^2/df = 1.062$, CFI = .999, RMSEA = 0 (90% C.I. = 0 – 1.0). This model fit the data equally as well as the unconstrained model, $\Delta\chi^2 = 3.609$, $\Delta df = 2$, $p = .165$, ΔCFI = .001, indicating metric invariance between the two groups. When measurement intercepts were constrained to be equal to test for scalar invariance, the fit of this model (Model C2) to the data was poor, $\chi^2(8) = 66.778$, $p < .001$, $\chi^2/df = 8.347$, CFI = .668, RMSEA = .175 (90% C.I. = .138- .215). A significant difference was found between this model and the measurement model, $\Delta\chi^2 = 62.528$, $\Delta df = 4$ $p < .001$, CFI difference = .331,

indicating that the indicator means are not equal between the typically-developing and clinical groups. Testing for partial scalar invariance could not be done with the two-factor model because there are only two indicators per factor and establishing partial scalar invariance requires that at least two indicators are constrained to be invariant (Byrne, Shavelson, & Muthén, 1989).

Differences between the groups. Table 7 shows unstandardized values of measurement parameters (factor loadings, subtest intercepts, and error variances). A single parameter value appears when a parameter was invariant across both groups. Different parameter values are presented when a parameter varied significantly across groups. In the two-factor model, the factor loadings were invariant across groups, indicative of metric invariance. However, all of the subtest intercepts varied between groups, with the clinical group's subtest intercepts systematically lower than those of the typically-developing group's intercepts. All of the error variances were also different between groups, though these differences did not appear to be systematic (two were higher and two were lower in each group).

In the one-factor model, the GWH, VRZ, and OIO factor loadings were invariant, indicating that the two groups do not use the same metric on the WHM subtest; the nonverbal construct does not appear to have the same meaning in both groups. Again, the item intercepts were not invariant between groups and the TBI group's were systematically lower than the typically-developing group's intercepts. Testing for the invariance of the error variances could not be completed, but did not appear to systematically differ between groups.

Table 7

*Invariant and non-invariant factor loadings, item intercepts, and error variances across 2 groups**

Latent variable	Subtest	Factor Loadings		Item Intercepts		Error Variances	
		TD	TBI	TD	TBI	TD	TBI
g	GWH	.921		56.214	45.815	24.47	20.166
	VRZ	1.086		55.273	44.093	51.106	15.215
	OIO	.502		57.46	49.444	57.123	85.635
	WHM	.449	1.225	54.107	49.130	116.13	51.468
Verbal	GWH	.931		56.214	45.815	23.242	20.02
	VRZ	1.074		55.273	44.093	51.821	14.885
Nonverbal	OIO	.666		57.46	49.444	55.405	81.109
	WHM	1.502		54.107	49.130	107.738	41.862

* Comparisons between traumatic brain injured (TBI) sub-sample and typically-developing (TD) group

GWH = Guess What; VRZ = Verbal Reasoning; OIO = Odd-Item-Out; WHM = What's Missing

RIAS and WISC-IV comparisons

Data Checking

Missing data. Of the 77 individuals in the clinical group who were administered WISC-IVs, 30 of these were administered incomplete WISC-IV tests (mean age = 12.2, *SD* = 2.98; 14 females; mean FSIQ = 86.7, *SD* = 19.3; 9, 2, 15, 1, and 3 from diagnostic categories 1-5, respectively). Data were missing because, in this clinical setting, select WISC-IV subtests may be administered without administering the entire test, depending on the reason for referral. Data from individuals who were not administered the full WISC-IV were excluded since predictive data replacement methods were unlikely to have produced meaningful scores for this highly heterogeneous clinical group. Demographics of the remaining 47 participants are: mean age = 12.15, *SD* =2.84 for WISC-IV and mean age =12.14, *SD* = 2.82 for RIAS; 20 female; 14, 2, 22, 2, 5, and 2 from diagnostic categories 1-6, respectively. See Tables 8 and 9 for zero-order correlations and descriptive statistics of RIAS and WISC-IV scores.

Age differences between testing. Seven participants were given the RIAS and WISC-IV at different times, with administration times ranging from one month to six months apart (*M* = 3.3 months). Four of these participants received the WISC-IV before the RIAS while for the remaining three, the administrations were reversed. RIAS *T*-score and index score differences were found for five participants when scores were calculated using the normative data for the age at which the WISC-IV was administered. None of the *T*-scores changed more than 1 standard point and, while one individual would have had a four point increase in CIX (89 to 93), the other index scores did not change more

than one or two points. These values are all contained within the 90% confidence interval around any score, so no data were excluded from this group.

Outliers, linearity, and homoscedasticity. Frequency tables and histograms were examined to identify possible univariate outliers in the distributions of RIAS *T*-scores, WISC-IV scaled scores, and index scores. Three univariate outliers were identified, as follows: 1) Participant A, 11.7 years old (diagnostic category 4), GWH = 28, OIO = 20, VRZ = 33, WHM = 9, VIX = 72, NIX = 41, CIX = 61, VCI = 45, PRI = 45, WMI = 62, PSI = 56, FSIQ = 42; 2) Participant B, 7.7 years old (diagnostic category 3), GWH = 74, OIO = 55, VRZ = 68, WHM = 68, VIX = 138, NIX = 122, CIX = 134, VCI = 128, PRI = 120, WMI = 100, PSI = 78, FSIQ = 112; and 3) Participant C, 11.3 years old (diagnostic category 3), GWH = 23, OIO = 24, VRZ = 32, WHM = 13, VIX = 66, NIX = 48, CIX = 49, VCI = 45, PRI = 45, WMI = 52, PSI = 53, FSIQ = 40. When these three outliers were removed, there was no more than a seven-point difference between any RIAS *T*-scores, no greater than a ten-point difference between RIAS index scores, no greater than a two-point difference between scores on each WISC-IV subtest (except for one participant who had a Digit Span score of 14, three standard score points higher than the next highest score), and for each WISC-IV index score, there was no difference larger than ten points. Since no remaining scores were greater than one standard deviation from the next lowest or highest score, no further univariate outliers were identified. The three univariate outliers were revealed to also be bivariate outliers through the examination of bivariate scatterplots between NIX and VIX, VIX and VCI, NIX and PRI, and CIX and FSIQ. Participant A's 31-point split between his/her NIX (41) and VIX (72) was the largest in the sub-sample. This was especially problematic

because no such split existed between his/her VCI and PRI (both 45), resulting in a CIX-FSIQ split of 19 points. Participant B showed significant discrepancies between his/her RIAS and WISC-IV scores, with a notable 22 point difference between the CIX (134) and the FSIQ (112). Participant C had a 21 point split between his/her VIX (66) and VCI (45) and a 9-point difference between his/her CIX (49) and FSIQ (40), a large difference at the tail of the distribution. The clinical implications of these discrepancies will be examined in the discussion section. No further bivariate outliers were identified. Examination of bivariate scatterplots also revealed approximate linearity and homoscedasticity between each pair of variables.

Assessment of normality. See Tables 8 and 9 for descriptive statistics. After removal of 30 cases with incomplete WISC-IV tests and three outliers (described above), only the Letter-Number Sequencing scaled score distributions was significantly skewed, $z = -2.55$, $p < .05$. No WISC-IV scaled score distributions had significant kurtosis. Only the OIO *T*-score distribution was significantly negatively skewed, $z = -2.80$, $p < .05$ and was significantly kurtotic, $z = 2.34$, $p < .05$. It should be noted that none of these values are significant when evaluated at an alpha-level of .01, as suggested by Tabachnick and Fidell (2007). Since the absolute values were not very high, and since all standard and scaled scores are in a meaningful metric, no transformations were made to the data prior to subsequent analyses. For RIAS scores in this sub-sample, Mardia's coefficient of kurtosis was 1.36, critical ratio = .651, indicating multivariate normality. The WISC-IV scores' distribution in the sub-sample also had multivariate normality, Mardia's coefficient of kurtosis = 2.3, critical ratio = .492. Informal comparison of the Q-Q probability plots of the RIAS *T*-scores and index scores with those of the WISC-IV

subtests and index scores revealed that the RIAS scores departed from normality even while the WISC-IV scores were spectacularly, almost creepily, normally distributed.

Table 8

Descriptive statistics of clinical sample's RIAS scores who had complete WISC-IVs

	n	*M*	S.D.	Skewness	S.E. Skew.	Kurtosis	S.E. Kurtosis
NIX	44	97.0	11.5	.047	.357	.427	.702
VIX	44	93.1	10.6	.153	.357	.054	.702
CIX	44	93.7	9.9	.378	.357	.856	.702
VCI	44	87.8	13.4	-.020	.357	-.615	.702
PRI	44	88.3	13.3	-.419	.357	-.420	.702
WMI	44	86.9	12.7	-.494	.357	-.348	.702
PSI	44	86.5	13.8	-.136	.357	-.299	.702
FSIQ	44	84.3	13.3	-.060	.357	-.671	.702

*$p < .05$

Note: Only data from complete WISC-IVs and corresponding RIASs are shown. Values rounded to the nearest tenth. NIX = Nonverbal Intelligence Index; VIX = Verbal Intelligence Index; CIX = Composite Intelligence Index; VCI = Verbal Comprehension Index; PRI = Perceptual Reasoning Index; WMI = Working Memory Index; PSI = Processing Speed Index; FSIQ = Full Scale Intelligence Quotient

Table 9

Descriptive statistics of clinical sub-sample's WISC-IV and RIAS subtest scores

	n	*M*	S.D.	Skewness	S.E. Skew.	Kurtosis	S.E. Kurtosis
BDN	44	7.89	3.04	-.131	.357	-.493	.702
SI	44	7.93	2.83	-.558	.357	.393	.702
DS	44	7.43	2.36	.335	.357	.330	.702
PC	44	8.66	3.03	-.533	.357	-.088	.702
CD	44	7.02	2.82	-.251	.357	-.824	.702
VC	44	8.07	3.02	.551	.357	.127	.702
LNS	44	8.34	2.94	-.909*	.357	.342	.702
MR	44	8.16	2.21	-.089	.357	-.743	.702
CO	44	7.98	2.68	-.536	.357	.237	.702
SS	44	8.09	3.12	-.180	.357	-.123	.702
GWH	44	44.9	7.1	-.003	.357	.360	.702
VRZ	44	43.4	9.3	.305	.357	.520	.702
OIO	44	49.9	6.9	-1.001*	.357	1.645*	.702
WHM	44	44.8	10.6	.349	.357	-.616	.702

* $p < .05$. Note: Only data from complete WISC-IVs and corresponding RIASs are shown. Values rounded to the nearest tenth. Subtest scaled scores (WISC-IV) and *T*-scores (RIAS) are shown. BDN = Block Design; SI = Similarities; DS = Digit Span; PC = Picture Concepts; CD = Coding; VC = Vocabulary; LNS = Letter-Number Sequencing; MR = Matrix Reasoning; CO = Comprehension; SS = Symbol Search; GWH = Guess What; VRZ = Verbal Reasoning; OIO = Odd-Item-Out; WHM = What's Missing

Index and Subtest Comparisons

See Table 10 for descriptive statistics and zero-order correlations between index scores. As predicted, the RIAS VIX was strongly correlated with the WISC-IV VCI, r = .715, $p < .01$, and had only a moderate correlation with PRI scores, r = .525, $p < .01$. Similarly, the NIX had a weak correlation with the VCI, r = .392, $p < .01$, but was only moderately correlated with PRI scores, r = .565, $p < .01$, which was equivalent to the correlation between VIX and PRI, $\chi^2 = 0.066$, *ns*. The VIX was moderately correlated with the Working Memory Index (WMI), r = .525, $p < .001$ while there was a low correlation between WMI and NIX, r = .328, $p < .05$. Both VIX and NIX were moderately correlated with the Processing Speed Index, r = .411 and r = .492, respectively, $p < .01$.

Paired-sample *t*-tests revealed significant differences between each pair of conceptually similar index scores from the RIAS and WISC-IV, as follows: VIX-VCI (mean difference = 5.32, *SE* = 1.42, 95% C.I. = 2.45-8.19), $t(43) = 3.74$, $p < .01$; NIX-PRI (mean difference = 8.73, *SE* = 1.76, 95% C.I. = 5.19-12.27), $t(43) = 4.97$, $p < .001$; CIX-FSIQ (mean difference = 9.41, *SE* = 1.30, 95% C.I. = 6.8-12.02), $t(43) = 7.26$, $p < .001$. In all cases, the RIAS index scores were higher than the WISC-IV index scores.

Table 10

Descriptive statistics of clinical sub-sample's RIAS and WISC-IV index scores

	n	*M*	S.D.	Skewness (*S.E.*)	Kurtosis (*S.E.*)	1	2	3	4	5	6	7
1. VIX	44	93.1	10.6	.153 (.357)	.054 (.702)	1						
2. NIX	44	97.0	11.5	.047 (.357)	.427 (.702)	.378*	1					
3. CIX	44	93.7	9.9	.378 (.357)	.856 (.702)	.828*	.828*	1				
4. VCI	44	87.8	13.4	-.020 (.357)	-.615 (.702)	.715*	.392*	.676*	1			
5. PRI	44	88.3	13.3	-.419 (.357)	-.420 (.702)	.525*	.565*	.647*	.534*	1		
6. WMI	44	86.9	12.7	-.494 (.357)	-.348 (.702)	.525*	.328*	.507*	.522*	.485*	1	
7. PSI	44	86.5	13.8	-.136 (.357)	-.299 (.702)	.411*	.492*	.545*	.409*	.630*	.448*	1
8. FSIQ	44	84.3	13.3	-.060 (.357)	-.671 (.702)	.701*	.566*	.763*	.808*	.843*	.732*	.775*

*$p < .05$

Note: Only data from complete WISC-IVs and corresponding RIASs are shown. Values rounded to the nearest tenth. NIX = Nonverbal Intelligence Index; VIX = Verbal Intelligence Index; CIX = Composite Intelligence Index; VCI = Verbal Comprehension Index; PRI = Perceptual Reasoning Index; WMI = Working Memory Index; PSI = Processing Speed Index; FSIQ = Full Scale Intelligence Quotient

Pearson's correlation coefficients were computed between RIAS subtest scaled scores and RIAS index scores and WISC-IV subtest standard scores (see Table 11). In this sub-sample, only the GWH score was significantly correlated with the other RIAS subtests: VRZ, $r = .54$, OIO, $r = .38$, and WHM, $r = .325$, all p-values $< .05$. OIO, WHM, and VRZ were not significantly correlated with each other. Similarly, GWH was significantly correlated with all WISC-IV subtests except Coding, $r = .293$, *ns*. Consistent with its conceptualization as a verbal, "*g*-saturated" test (Reynolds & Kamphaus, 2003), GWH's highest correlations were with verbal and fluid reasoning WISC-IV subtests, with a strong correlation with Vocabulary, $r = .775$, and moderate correlations with Similarities, $r = .566$, Matrix Reasoning, $r = .531$, and Comprehension, $r = .509$. GWH was also moderately related to Symbol Search, $r = .457$, and Picture Concepts, $r = .428$, reflecting the conceptualization of Picture Concepts as a verbally-mediated task. Finally, GWH had low correlations with the visuo-spatial task, Block Design, $r = .377$, and the working memory tasks, Letter-Number Sequencing, $r = .357$, and Digit Span, .314, all p-values $< .05$.

The Verbal Reasoning subtest was moderately correlated with Digit Span, $r = .636$, Vocabulary, $r = .537$, and Similarities, $r = .525$ while having low correlations with Comprehension, $r = .399$, Picture Concepts, $r = .373$, Symbol Search, $r = .338$, and Block Design, $r = .334$, all p-values $< .05$. VRZ was not significantly correlated with Coding, Letter-Number Sequencing, or Matrix Reasoning.

The Odd-Item-Out subtest was moderately correlated with the verbal subtests, Similarities, $r = .467$ and Vocabulary, $r = .451$, with the visuo-spatial and fluid reasoning tasks, Block Design, $r = .402$ and Matrix Reasoning, $r = .401$, and with the psychomotor

speed task, Symbol Search, $r = .404$, all p-values $< .05$. Low correlations were also found between OIO and the verbal subtest, Comprehension, $r = .350$, and with Picture Concepts, $r = .350$ and the two working memory subtests, Letter-Number Sequencing, $r = .364$ and Digit Span, $r = .312$. Despite its moderate correlation with Symbol Search, OIO was not significantly correlated with Coding.

Consistent with expectations, What's Missing was significantly correlated with the two "performance" subtests, Block Design, $r = .383$ and Matrix Reasoning, $r = .307$, both p-values $< .05$, though these were small correlations. What's Missing also had low correlations with Coding, $r = .38$ and Vocabulary, $r = .32$, $p < .05$. It was not significantly correlated with Picture Concepts, Similarities, Comprehension, Digit Span, Letter-Number Sequencing, or Symbol Search.

Table 11

Zero order correlations between RIAS and WISC-IV subtest standard scores

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. GWH	1													
2. VRZ	.539*	1												
3. OIO	.38*	.289	1											
4. WHM	.325*	.109	.198	1										
5. BDN	.377*	.334*	.402*	.383*	1									
6. SI	.566*	.525*	.467*	.088	.317*	1								
7. DS	.314*	.636*	.312*	.069	.117	.395*	1							
8. PCn	.428*	.373*	.35*	.269	.478*	.448*	.269	1						
9. CD	.293	.13	.283	.38*	.541*	.151	.257	.30*	1					
10. VC	.775*	.537*	.451*	.32*	.233	.577*	.378*	.575*	.226	1				
11. LNS	.357*	.283	.364*	.174	.301*	.218	.351*	.411*	.321*	.411*	1			
12. MR	.531*	.274	.401*	.307*	.480*	.481*	.259	.457*	.544*	.486*	.525*	1		
13. CO	.509*	.399*	.35*	.091	.171	.475*	.41*	.486*	.341*	.669*	.441*	.264	1	
14. SS	.457*	.338*	.404*	.291	.511*	.388*	.286	.39*	.467*	.454*	.408*	.403*	.357*	1

*$p < .05$ Note: Values rounded to the nearest tenth.

Discussion

The purpose of the current study was to examine whether the RIAS fit the CHC and Cattell-Horn G*f*-G*c* theories of intelligence upon which it was based and, in particular, to assess the validity of the RIAS' index and subtest scores. This was accomplished through comparison of factor structures between typically-developing and mixed clinical groups and between scores on the WISC-IV and the RIAS in the clinical group. The results were mixed, with some findings supporting the validity of the CIX, NIX, and VIX while others suggested that the CIX and NIX in particular should be interpreted with caution, depending on the population in which the RIAS is being used. The functioning of OIO and WHM in all analyses suggested that the RIAS' nonverbal subtests are most problematic with the greatest impact on the RIAS' validity as a measure of overall, verbal, and nonverbal intelligence.

The one- and two-factor models fit the data of the typically-developing group equally well, although the two-factor model may have been slightly more parsimonious than the one-factor model in this group. While the two-factor model also fit the clinical group's data well, the fit of the one-factor model was equivocal; the chi-square and CFI indicated a good fit while the RMSEA indicated a poor fit. However, the RMSEA can be a misleading fit index when the sample size is not large and the degrees of freedom are small (Kenny, 2010) as was true in the current study. Given the large 90% confidence interval of the RMSEA value, it seems that the greater amount of variability in the clinical group may have been partially to blame for these conflicting fit indices. Indeed, there was a wider range of scores in the clinical group and a large proportion of this group had large, significant splits between their nonverbal and verbal subtest scores.

While this would have enhanced the fit of the two-factor model in the clinical group, it likely decreased the fit of the one-factor model in this sample. Nevertheless, the one-factor model should not be rejected outright as a bad fitting model in the clinical group for the following, additional reasons: a) most of the fit indices indicated that the one-factor model fit the data of the clinical sample; b) the factor loadings of all subtests were high on the general factor; c) the RMSEA's confidence interval extended down to zero; and d) the PCLOSE indicated that the RMSEA was likely a value that would indicate good fit.

The current study's results are simultaneously consistent and inconsistent with the findings of previous studies that have examined the factor structure of the RIAS. In general, past studies using CFA have found evidence for the interpretability of the NIX and VIX through the good fit of a two-factor (verbal/nonverbal) model to RIAS data in both typically-developing individuals (Reynolds & Kamphaus, 2003) and two referred samples (Beaujean et al., 2009), congruent with the current study's findings. However, in contrast to the current study's findings, neither of these CFA studies could provide evidence of the validity of the CIX since the one-factor model did not fit the data of any of their samples. On the other hand, previous studies that utilized EFA techniques argued for the validity of the CIX and cautioned against interpreting the NIX and VIX when their methods found evidence of a strong, general factor emerging from RIAS data of referred students (Nelson et al., 2007) and of the RIAS' normative sample (Dombrowski et al., 2009). The current study is the first to find equal fit of one- and two-factor models to the RIAS data of a typically-developing sample, and so to provide evidence for the validity of the NIX, VIX, *and* the CIX using a single factor analysis technique. However,

there are several reasons to be cautious about concluding that all three RIAS indexes should be interpreted, as outlined below.

Although previous authors (e.g. Nelson et al., 2007; Dombrowski et al., 2009) have argued that it is more defensible to interpret the CIX than the VIX and NIX, the results of the current study do not unequivocally support this contention. To wit, the fit of the one-factor model to the clinical sample's data was bad according to the RMSEA. This bad fit was found despite the uniformly high loadings of each of the subtests on the general factor and the very high correlation (r = .90) of the verbal and nonverbal factor in the two-factor model. However, as mentioned previously, the RMSEA may not be a good indication of model fit in this sample because of the small number of degrees of freedom and the relatively small sample size. In fact, if the sample size of the clinical group had been equal to that of the typically-developing group (n = 187), an increase of only 23 participants, the RMSEA would have fallen into the adequate fit range, assuming no substantial increase in the chi-square value. On balance, there is stronger evidence that the one-factor model provided good fit rather than bad fit to the clinical sample's data, but since the RIAS is intended for decision-making purposes in a clinical population (Reynolds & Kamphaus, 2003), caution may still be warranted in interpreting the CIX in light of the equivocal CFA results.

The murkiness of the meaning of the RIAS index scores becomes apparent when trying to consider why the one-factor and two-factor models fit the typically-developing sample's data equally. There is strong evidence that crystallized and fluid abilities separately explain a substantial proportion of variance in various outcome measures (e.g. specific academic abilities) over and above *g* (e.g. McGrew, Flanagan, Keith, &

Vanderwood, 1997; Vanderwood, McGrew, Flanagan, & Keith, 2002). Since these factors are separable from each other and from *g*, the very high correlations between the nonverbal and verbal factors in both samples, and the equal fit of the one- and two-factor models in the typically-developing sample (and likely the clinical sample) brings into question the validity of the NIX and VIX. Certainly, it is difficult to argue that two different aspects of intelligence (verbal/crystallized and nonverbal/fluid) are being measured by the RIAS given the large proportion of variance shared by these factors in the current and past studies (e.g. Reynolds & Kamphaus, 2003; Beaujean et al., 2009). This pattern has also been mirrored in studies that used principal factors analysis with both orthogonal and oblique rotations. That is, although the nonverbal and verbal subtests loaded on two theoretically consistent factors, the factors were highly correlated (r = .79 in both Nelson et al., 2007 and Dombrowski et al., 2009). In fact, the initial *g*-loadings of all subtests in these studies were so high that exploration of further factors (i.e. rotation) would not have been warranted without *a priori* reasons to do so. Finally, after comparing RIAS index scores with the Woodcock-Johnson-III-Cognitive (WJ-III-Cog) test's General Intellectual Ability Index, Gc, and Gf indexes, at least one previous study (Krach, Loe, Jones, & Farrally, 2009) concluded that, while crystallized intelligence and even general intellectual ability are adequately measured by the VIX and CIX, fluid intelligence is not measured by the NIX. Combined with the results of these past studies, the current findings indicate that the verbal and nonverbal factors may not be substantively different enough to justify interpreting the NIX and VIX separately and meaningfully (Dombrowski et al., 2009; Nelson et al., 2007).

As alluded to, the NIX may be particularly problematic given the variable factor loadings of OIO and WHM. As has been found previously (Reynolds & Kamphaus, 2003; Schraw, 2005; Dombrowski et al., 2009; Nelson et al., 2007), GWH and VRZ had consistent, high *g*-loadings in both models and both samples in the current study. However, OIO and WHM had much weaker loadings on the nonverbal and *g* factors than the two verbal subtests, at least in the typically-developing sample. Previous studies have found that the RIAS's nonverbal subtests "behave" in a much less stable manner than the verbal subtests. Using either CFA (Reynolds & Kamphaus, 2003) or EFA methods (Dombrowski et al., 2009; Nelson et al., 2007), these subtests do not load as highly with each other as the verbal subtests do. The technical manual (2003) reported that WHM even loaded with verbal subtests in the 12-18 year old group while loading with nonverbal subtests (OIO and nonverbal memory) in other age groups in the normative sample.

Nonverbal Subtests

Of particular concern are the low loadings of the WHM subtest onto both factors in the typically-developing sample. What's Missing is essentially identical to the Picture Completion subtest of the Wechsler Intelligence Scales for Children – Third Edition (WISC-III; Wechsler, 1996) which had a *g*-loading of .60 and was subsequently moved to the supplementary battery on the WISC-IV. It seems that WHM is operating even more poorly on the RIAS. Indeed, although WHM loads significantly onto both the *g* and nonverbal factors, neither factor accounts for a substantial proportion of variance in this subtest (7% and 11%, respectively). Inspection of the zero-order correlations of the subtests reveals that WHM shares little variance with the other subtests, including OIO.

In fact, it was equally as correlated with the verbal subtests as it was with this other "nonverbal" subtest in both samples. This would suggest that a different factor or factors may drive scores on the WHM subtest than the one(s) that underlie scores on the other three subtests of the RIAS.

Examination of the WHM subtest may give some clue as to why it did not load on the same factors with the other RIAS subtests. The Technical Manual (Reynolds & Kamphaus, 2003) describes the WHM subtest as follows:

> What's Missing requires the examinee to apply spatial-organizational and visual-perceptual skills to develop an organized scheme, or gestalt, of the picture presented. The examinee must first synthesize the elements of the picture, and then, through visual analysis, discover the essential missing element of the picture… Little prior specialized knowledge is required beyond the most rudimentary incidental knowledge of everyday life (e.g. knowing the basic elements of a face by sight, not by name). Because examinees may respond either verbally or nonverbally, the knowledge requirement is reduced even further. What's Missing thus requires the integration of visual-perceptual skills with spatial and non-verbal reasoning to deduce a solution. This task is therefore more closely aligned with nonverbal intelligence and fluid ability than with verbal intelligence and crystallized knowledge (p. 92).

Although the authors claim that performance on the WHM does not depend on acquired knowledge, all of the items do in fact *require* at least some prior experience with the stimuli presented. For example, an examinee might be presented with a picture of an American penny with the date missing. In order to respond correctly to this item, they would need experience handling a penny or seeing it up close and attending to it. This

characteristic of WHM seems to be inconsistent with the conceptualization of fluid intelligence as the ability to reason when presented with novel tasks or using over-learned elements (e.g. letters of the alphabet; Horn & Cattell, 1966). In fact, intuition would suggest that performance on WHM is more dependent on crystallized intelligence. However, the low loadings on a common factor with GWH and VRZ would be surprising if WHM were tapping crystallized ability as these two subtests likely do. Thus, it is difficult to characterize what construct or aspect of intelligence WHM is measuring, or even if it should be considered to be providing information predominantly about intellectual ability. Instead, performance on WHM may depend more on visual scanning ability, processing of local versus global features, and/or acquired knowledge. These possibilities will be explored below when the RIAS-WISC-IV comparisons are discussed.

The low and moderate loadings of the OIO subtest on *g* and the nonverbal factors, respectively, in the typically-developing group also have implications for the interpretability of the NIX and CIX. This finding is congruent with the results of EFA (Dombrowski et al., 2009; Nelson et al., 2007; Reynolds & Kamphaus, 2003) and CFA studies (Reynolds & Kamphaus, 2003; Beaujean et al., 2009) that found OIO generally loaded lower than the verbal subtests on *g*. In these studies, OIO loaded on the nonverbal factor with greater or equal magnitude as the WHM subtest in the 3-18 year olds, but these loadings were not as high as the verbal subtests' loadings on the verbal factor. Concerns with OIO's cross-loadings with VRZ and GWH have been raised previously (Dombrowski, 2008; Nelson et al., 2007; Dombrowski et al., 2009). As noted by the test authors (Reynolds & Kamphaus, 2009), verbal strategies can be used to solve these nonverbal subtests (they even recommend asking the examinee how they solved the tasks

if use of verbal strategies is suspected). The OIO subtest may load with VRZ and GWH because performance on it involves (perhaps inconsistently) some verbal abilities. Its loading with these verbal tasks may be lowered by the fact that OIO is presented in a visual format while both subtests are presented in an auditory modality. On the other hand, OIO and WHM have a common presentation format and so both require at least some visual and visual-spatial abilities, which would partially account for some of their shared variance. The OIO subtest's correlations with the WISC-IV Block Design and Matrix Reasoning subtests are testament to its visual-spatial components while its equal correlations with Similarities and Vocabulary support the contention that performance on OIO is, at least in part, verbally-mediated. Of note, given its barely moderate correlation (r = .40) with Matrix Reasoning, OIO is likely not just a "reverse matrix analyses task" (Brueggemann, Reynolds & Kamphaus, 2006, p. 133).

The mixed nature of OIO may also extend to the item level. For example, on the first 12 items of OIO, the odd item out is strongly perceptually different from the others (e.g. different colour, different simple geometric shape). Thereafter, there is a mix of items where solutions can be made through much more overt and verbal higher-order classification (e.g. living animals versus a stuffed animal) with yet other items of abstract drawings that could be solved verbally or via visual comparison of lines and angles among items. That is, there seems to be a choice to solve the OIO items either verbally or nonverbally, or to use a mixture of the two strategies. However, this subtest had high internal consistency coefficients across all age groups (Reynolds & Kamphaus, 2003), indicating that one predominant strategy tended to be used on all items by each individual examinee.

Clinical versus Typically-Developing Group

In the clinical sample, the verbal subtests once again loaded very highly onto all factors, with factor loadings remaining essentially unchanged for GWH and VRZ in both models. However, in both the one- and two-factor models, OIO and WHM had high factor loadings on the nonverbal and *g* factors. In this context, it is unsurprising that the verbal and nonverbal factors were so highly correlated (r = .90). Despite the high factor loadings of all subtests on the general factor, and in contrast to the results in the typically-developing sample, the one-factor model fit indices were not unanimously good. The reason for why only the two-factor model fit the clinical group's data well while both the one- and two-factor models fit the data of the typically-developing sample may lie with: i) the operation of the nonverbal subtests in these models; ii) the age range included in the current study; iii) the consistently high socioeconomic status and environmental homogeneity of the samples; and iv) how the measurement of *g*, G*f*, and G*c* might be affected in typically-developing groups versus those samples comprised of individuals with various developmental disorders and neurological insults.

The RIAS subtests were designed to be "*g*-saturated." Why then, does the two-factor model have a better fit than the one-factor model in the clinical population and why do the g-loadings of the RIAS's subtests vary according to the population tested? The high factor loadings of all subtests on their respective factors, coupled with the very high correlation between the factors themselves in the clinical group may indicate a kind of paradox arising from testing these models in such a heterogeneous clinical sample. That is, some neurological insults may lead to a global impact on *g*, thus impacting scores on many subtests. Other insults may have differentially affected various aspects of brain

integrity that impact cognitive abilities underlying *g*, G*c*, and G*f*, etc., leading to a number of individuals with substantial divides between their performances on VRZ and GWH versus on OIO and WHM. The latter cases would cause a poor fit of the one-factor model to the sample's data while the former cases would lead to a correlation between the nonverbal and verbal factors and high loadings of all subtests on their respective factors since the factors were allowed to covary. This idea will be explored further below, beginning with a discussion of the nature of *g*.

As outlined previously, the "*g* factor" refers to the component variance common to all cognitive tests (Spearman, 1904; Jensen, 1998). Tests with higher *g*-loadings tend to involve complex cognitive operations (e.g. abstraction, and inductive and deductive reasoning) while tests with lower *g*-loadings rely on less complex cognitive abilities (e.g. rote memory, simple reaction time, sensory discrimination; Larson, Merritt, & Williams, 1988). As such, it is thought that *g* is psychometrically, but not physiologically (at the brain level), unitary (Gottfredson & Saklofske, 2009). That is, *g* is a reflection of a complex system that is itself comprised of a number of basic cognitive abilities working to produce some output (Detterman, 1987, 1994, 2002). It is not a direct measure of knowledge or skills, but of the capacity and efficiency of processes by which these things are acquired and applied (Jensen, 1998). Put another way, *g* is not an entity or ability itself, but is the manifestation of the action of a number of brain properties that produce a wide number of cognitive abilities (Jensen, in Flynn, 2000; Gottfredson & Saklofske, 2009). This explains why *g* emerges psychometrically from tests reliant on these cognitive abilities, regardless of what the tests are. On the other hand, it has been

suggested that language and spatial visualization are more localized in the brain, allowing their psychometric factors to vary more independently (Gottfredson & Saklofske, 2009).

While some theorists have argued that the lion's share of variability in *g* is accounted for by genetic factors (Jensen, 1998; Herrnstein & Murray, 1994), the roles of environmental factors and reciprocal gene-environment interactions in influencing *g* (and its proxy, IQ) has been established (e.g. Dickens & Flynn, 2001). These factors shape the brain's physical and functional features, including: glucose metabolism; dendrite length; volume of whole and individual parts; amplitude and latency of resting and evoked electrical potentials (Gottfredson & Saklofske, 2009); degree of axonal myelination; neural signalling efficiency; and the total number of neurons in the brain (Jensen, 1998). How and to what degree these features develop and remain intact influence the development and operation of various cognitive abilities. If the brain develops atypically or incurs an insult, the operation of some or all of these cognitive abilities will also be impacted (Detterman, 1987; Schneider, 2011). This would affect the intercorrelations among tests (including the *g* component) relying to various degrees on differing underlying abilities. It would also affect performance on tests that measure products (e.g. knowledge) whose acquisition is dependent on these abilities and which are often more difficult to acquire when the abilities are undermined (i.e. the "Matthew effect"; Stanovich, 1986; Shaywitz et al., 1995; Dennis et al., 2009). The extent to which abilities are co-affected would theoretically be related to how widespread or focal the acquired damage was or, in atypically-developing groups, whether atypical brain development affected cognitive abilities globally (e.g. Down's Syndrome) or affected only some abilities while leaving others relatively spared (e.g. William's Syndrome). A deleterious

genetic or environmental factor (e.g. lead poisoning in a child) impacting all cognitive abilities would be expected to produce higher intercorrelations among subtests and thus, "higher" *g* in these individuals than in those with more typical development or without such exposure. On the other hand, a neurological insult can produce global or focal damage, and could even produce both as in a child who "grows into" his/her widespread deficits after an injury that had initially appeared to impact only a few abilities (Dennis, 2000). In theory, depending on its location, more focal damage, or damage that occurred later in development might impact only a few cognitive abilities while leaving others relatively intact. In such a case the intercorrelations among cognitive tests relying to various extents on each of the spared and damaged abilities would be lowered, thus producing a "smaller" *g*. In this scenario, two or more factors could emerge as a result of this fractionation.

The current study's mixed clinical sample included individuals with a large variety of disorders, disease processes, and acquired injuries. There was also a wide range in terms of when the various insults were incurred, and so the degree to which typical brain development occurred, if at all, varied across individuals in this sample. It would be unlikely that all individuals in the sample experienced either focal or global deleterious effects on their brains, but the fact that a two-factor model fit the data better than a one-factor (*g*) model suggests that, on balance, there were differential effects on cognitive abilities contributing to verbal/auditory versus nonverbal/visual-spatial performances in this sample. It is interesting that the loadings between the two verbal subtests remained essentially unchanged between the two models; the improved loadings of the two nonverbal subtests were what underlay the improved fit indices of the two-

factor model compared to the one-factor model. This may be reflective of GWH and VRZ being heavily reliant on vocabulary, something that remains relatively intact after an acquired brain injury (Lezak, Howieson, & Loring, 2004), which was a highly represented diagnostic category in the clinical sample. Ultimately, the unanimously high loadings of all of the subtests on *g* suggests that, in this sample of developing children, all cognitive abilities were impacted to some degree.

Indeed, the very high correlation between the Verbal and Nonverbal factors in this sample may be reflective of the pervasive effects of any neurological insult on the developing brain. As well, even in the one-factor model, both nonverbal subtests load more highly on *g* in the clinical sample than in the typically-developing sample. The low loadings of WHM on *g* in the typically-developing sample indicate that this subtest is not a good measure of *g*; it may rely to a greater extent on less complex abilities such as visual acuity and visual scanning (as Picture Completion does; Sattler, 2008). The high loading of WHM on *g* in the clinical group indicates that, on balance, there was enough impairment of cognitive abilities (even less complex ones) in this sample that performance on all tests were affected to some degree, increasing the correlations among subtests. Alternatively, this result may be reflective of the heterogeneity in the clinical sample: all cognitive abilities of some individuals were impacted to approximately the same degree (leading to higher loadings on *g*) while others had various cognitive abilities affected to different degrees, leading to large splits in performance on GWH/VRZ versus OIO/WHM and decreasing the fit of the one-factor model.

However, this possibility does not account for the fact that both the clinical and typically-developing samples had large proportions of individuals with substantial

NIX/VIX splits. Indeed, 33.69% of the typically-developing sample had NIX-VIX differences of at least one standard deviation (7.5% of these were two *S.D.* apart) compared with 35.4% in the clinical sample with a one *S.D.* NIX-VIX difference and 3.7% with a two *S.D.* difference. Furthermore, since nine to ten points between the NIX and VIX is statistically different (Reynolds & Kamphaus, 2003), 53.5% of the typically-developing sample had a significant difference between these index scores. Composite scores are more reliable than the subtest or index scores that comprise them (as per the Spearman-Brown prophecy; Kaplan & Saccuzzo, 2005). When significant differences exist between sub-index scores, it is not always advisable to interpret the composite score (Zhu & Weiss, 2005). If the majority of the typically-developing sample's NIX scores diverged significantly from their VIX scores making the CIX a less meaningful index, yet there is not good evidence for interpreting the NIX and VIX scores separately, psychologists may often be placed in a difficult spot trying to determine which index score to interpret. In fact, in such cases, none may be meaningful representations of intellectual ability.

The verbal subtests of the RIAS seem to be less problematic than the nonverbal subtests since they consistently have high loadings on both one- and two-factors in every study, both current and past. While much of the evidence in the current study strongly indicates that the two verbal RIAS subtests are verbal tasks that relied on crystallized abilities, OIO and WHM do not "hang together" very well in the factor analyses, are not related to all of the same WISC-IV subtests, and in fact, are not always related to the WISC-IV subtests with which they are theoretically meant to be most highly related. These subtests seem to be measuring a variety of abilities, some of which are verbal,

some nonverbal/spatial, and this "fuzzy" measurement may contribute to the lower loadings with a nonverbal factor and a common factor. An examination of the correlations between RIAS scores and WISC-IV scores may provide more information about what the RIAS indexes and subtests, especially the nonverbal ones, measure.

The Relationship between RIAS and WISC-IV Scores: Clues about What the RIAS Subtests Measure

Intercorrelations among the RIAS and WISC-IV index scores provided evidence for the validity of the CIX, but indicated some issues with the NIX and VIX. The moderately high correlation (r = .76) between the CIX and the WISC-IV Full-Scale IQ score is in accordance with recommendations for the relationship between composite scores of the *Standards for Educational and Psychological Testing* (*Standards*; AERA, 1999). In agreement with the CHC model underlying the RIAS, the VIX was appropriately highly correlated with the Verbal Comprehension Index, though its moderate correlation with the Perceptual Reasoning Index was perhaps too high. However, this is not necessarily problematic since the RIAS's subtests were designed to be *g*-saturated, which would allow for this overlap with nonverbal reasoning tasks, including, in particular, Matrix Reasoning. Both GWH and VRZ appear to be measures of verbal, crystallized abilities, given their relationships with all of the WISC-IV subtests. GWH in particular was related to almost all subtests, and so may be the most *g*-saturated RIAS subtest. The verbal subtests, especially VRZ, also appear to rely on auditory working memory ability, perhaps because the verbal subtests are presented orally to examinees (as are Digit Span and Letter-Number Sequencing). Consistent with this is the lower relationship between the Working Memory Index and the NIX, the latter index

being comprised of subtests with visually-presented stimuli. The low-moderate relationships between the Processing Speed Index and each RIAS index score are consistent with the authors' (Reynolds & Kamphaus, 2003) contention that the RIAS is not heavily reliant on psychomotor speed.

More problematic were the correlations with NIX. Though the NIX had an appropriately low correlation with the VCI, its correlation with the PRI was not as high as might be expected if both indexes are comprised of subtests tapping nonverbal, fluid reasoning. The RIAS authors (Reynolds & Kamphaus, 2003, 2009) have previously argued that the lower correlations between NIX and PRI compared to that between VIX and VCI is due to PRI's reliance on psychomotor speed and motor abilities (Reynolds & Kamphaus, 2003). However, OIO and WHM were equivalently correlated with Block Design, which has clear motor demands, and with Matrix Reasoning, which is a measure of "primarily fluid and novel reasoning" (Sattler, 2008, p.285) with no more motor requirements than those of OIO and WHM. These correlations were all fairly small, as was OIO's correlation with Picture Concepts and, in fact, WHM was unrelated to this so-called perceptual reasoning subtest at all.

Prior studies have indicated that Picture Concepts (PCn) is only a fair measure of *g*, that it contributes only moderately to the PRI, and that it is also dependent in part on crystallized knowledge and language ability (Sattler, 2008), something supported in the current study given its consistent correlations with the verbal subtests of both the RIAS and the WISC-IV. If PCn is a verbal task, OIO's relationship with it, along with its consistent, moderate correlations with Similarities, Vocabulary, Comprehension, Guess What, and Verbal Reasoning, suggest that it too is a task that relies, at least partially, on

verbal abilities in a clinical sample. It may be that OIO's correlations with verbal tasks would be higher if they were all presented in the same modality. Indeed, OIO's equivalent correlations with BDN and MR are unsurprising since all tests require scanning, comparison, and reasoning in a visual modality. If OIO is a verbally-mediated, visually-presented task, the cross-loadings between the nonverbal and verbal factors and the fit of the one-factor model are expectable. However, the misfit of the one- and two-factor models seems to be more clearly affected by the loadings of WHM, which is not surprising since it is not a highly *g*-saturated task (the evidence for this is also apparent with its kin, Picture Completion; Sattler, 2008) but also does not appear to measure any one intellectual ability in particular, as is apparent in its relationships with WISC-IV subtests.

As was clear through examination of factor loadings, the WHM subtest is least related to the other RIAS subtests, even to OIO. This may have been because of several reasons. Firstly, WHM was not related to either of the working memory subtests. Since working memory has been identified as particularly related to *g* (e.g. Colom et al., 2009), WHM's status as *g*-saturated cannot be secured. Secondly, while VRZ, GWH, and OIO appear to be more verbally-mediated, WHM seems less dependent on verbal abilities. Although WHM had a small correlation with GWH and Vocabulary, it is perhaps surprising that it was not more related to crystallized knowledge tasks. Intuition suggests that previous experience (and crystallized abilities) with some of the scenes and objects presented as WHM stimuli would be necessary for successful performance on this task. On the other hand, WHM was unrelated to Picture Concepts, which is surprising if both tasks are meant to assess nonverbal reasoning (G*f*; Keith, Fine, Reynolds, Taub, &

Kranzler, 2006) with pictures. Furthermore, with only low correlations with Matrix Reasoning and Block Design, it is not clear to what extent WHM relies on spatial reasoning abilities. In fact, given WHM's equivalently high relationship with Coding, it may be that variance shared between WHM and these spatial tasks is a result of all of the tasks requiring visual acuity, visual-sensory scanning, visual and verbal elaboration, perceptual discrimination, persistence, global and local feature processing, attention to detail, and the ability to focus and perform under time constraints (Kaufman and Lichtenberger, 2000; Flanagan and Kaufman, 2004; Kaufman, 1994; Sattler & Dumont, 2004).

How the RIAS Measures Intelligence in a Developing Population of Children

If performance on the nonverbal subtests, especially OIO, can be mediated verbally or visual-spatially, then another reason for the lower loadings of these subtests in both models could be that different strategies were used by different participants to solve items on these tests. This possibility is even more likely given the current study's samples of developing children, a population whose use of strategies in problem-solving is already highly variable. The higher variability in the factor structure of children's scores was shown in studies using the Schmid-Leiman procedure to examine the higher-order structure of the RIAS. In these studies, the verbal subtests were more cohesive, behaved more consistently, and more of their variance was accounted for than the nonverbal subtests (Dombrowski et al., 2009; Nelson et al., 2007). This was especially true in the 6 to 11 and 12 to 18 year old groups, which almost exactly match the age group of the current study. To illustrate, the communalities of OIO and WHM were relatively lower and the uniquenesses correspondingly higher than those of GWH and

VRZ, especially in the 6-11 and 12-18 year old groups (Dombrowski et al., 2009; Nelson et al., 2007). For example, in 12-18 year olds, the proportion of variance accounted for in WHM by the general and nonverbal factors was 29% with 71% of its variance accounted for by unique influences (Dombrowski et al., 2009). Furthermore, the variance accounted for in the nonverbal subtests by the general factor tended to be more variable in the 6-11 and 12-18 year olds than in other age groups. For example, *g* only accounted for 18% of the variance in OIO scores in 6-11 year olds and only 12% of variance in WHM in 12-18 year olds (compared to 27-47% of variance in these subtests in older and younger groups). These lower factor loadings of the nonverbal subtests in the 6-18 year olds reflect the results of the current study and suggest that OIO and WHM are especially unstable in this age range, at least in groups that are developing typically. This could reflect the quality and enormous quantity of cognitive development across this age range. Specifically, it has been shown that children's use of strategies in problem-solving shows a great deal of variability, not just over time, but also within the same session, across different but related problems, and even when solving the same problem twice in a single session (Siegler, 1994). If there are a greater number of strategies that can be used to solve the RIAS's nonverbal tasks than the verbal tasks (e.g. both verbal and nonverbal strategies), if the grasp and use of these various strategies develop at different rates across childhood, and if a child uses a mixture of these strategies across different items of a given subtest, this would lead to greater variability in performance on these subtests but decrease loadings with subtests (i.e. GHW and VRZ) more reliant on one or two cognitive abilities or with less problem-solving required (e.g. rehearsal and retrieval of verbal information). Consistent with this, when Beaujean et al. (2009) estimated the

reliabilities of the verbal and nonverbal factors in two separate samples of referred students, ages 6-18 years olds, they found that the verbal factors had uniformly high reliabilities (.82) in both samples while the nonverbal factor had consistently lower reliabilities (.58 and .63). The latter lower reliabilities are more likely related to developmental differences in use of cognitive strategies (i.e. a measurement issue in a developing sample) rather than differences in the structure of intelligence across development since the three-stratum factor structure of intelligence has been shown to be invariant over the lifespan (Bickley, Keith, & Wolfle, 1995).

The limited and young age range (4 – 18 years) included in the current study may be partially responsible for the divergence from the results of past CFA studies. Specifically, the results of Dombrowski et al.'s (2009) follow-up study that used the Schmid-Leiman procedure to examine the higher-order structure of the RIAS in the original normative sample hint at the verbal and nonverbal factors operating differently in groups of individuals ages 6-18 years than in younger (3-5 years) or older (19-94 years) groups. Firstly, this study concluded that there was not enough evidence to support the higher-order structure of the RIAS and argued against interpreting the NIX and VIX. In the total sample, the general factor accounted for 29% (WHM) to 46% (GWH) of the subtest variance and 38% of the total variance and 66.6% of the common variance across the age ranges while the verbal factor accounted for an additional 11.7% of the total variance and 20.4% of the common variance and the nonverbal factor accounted for 7.4% of the total variance and 12.9% of the common variance beyond the general factor (Dombrowski et al., 2009). However, compared to other age groups, in the 6-11 and 12-18 year olds, the first-order factors - especially the verbal factor – accounted for a greater

amount of common variance beyond that accounted for by *g*. For example, the verbal factor accounted for 28.2% and 27% of common variance beyond that accounted for by *g* in GWH and VRZ scores of 6-11 year olds and 12-18 year olds, respectively, compared to 19.5% and 17.2% of common variance in these scores of 3-5 year olds and 19-94 year olds, respectively. The common variance accounted for by the nonverbal factors ranged from 14.7%-15.6% in 3-18 year olds compared to 11.8% in 19-94 year olds, which was likely not a significant difference. These results suggest that there is a separate portion of variance not accounted for by a general factor which can be accounted for by an orthogonal verbal factor and which is especially apparent in children aged 6-18 years old. Indeed, when Reynolds & Kamphaus (2003) used CFA in this data set, they found that the two-factor, but not the one-factor, model fit the data for all age groups. This begs the question: Why did the one-factor model fit in the current study's samples but not in the RIAS' normative sample?

The Impact of Demographic Homogeneity on the Results

One reason might be that the normative sample necessarily included a diverse sample, stratified by ethnicity and region (Reynolds & Kamphaus, 2003) while the typically-developing sample was drawn from Victoria and the clinical sample was drawn from Vancouver Island. This would introduce a large degree of homogeneity into both of the current study's samples with respect to the physical environment (e.g. exposure to pollutants pre- and post-natally), quality of schooling (Bouchard & Segal, 1985; Ceci, 1991), cultural (Hunt, 1997), and other environmental variables that can impact scores on cognitive tests. While information on ethnicity of the clinical sample is not available, the majority in the typically-developing sample were White and it is believed that the clinical

sample was also mostly White. Furthermore, the socioeconomic status of these samples is likely to be consistently high since Victoria's median total income (British Columbia index) is 1.18, meaning that it is 18% higher than the province's median total income (BC Stats, 2008). As testament to this, over 84% of the children in the typically-developing sample had at least one parent with a post-secondary degree and for most children, both parents had *at least* one degree or diploma. Furthermore, of nine patients in one random sample of patients seen in the neuropsychology department at QAACH, five had parents with at least one post-secondary degree and all had parents who had at least finished high school, indicating that the clinical sample was also characterized by relatively high SES (M. Joschko, personal communication, August 10, 2011). The socioeconomic status of children's parents impacts both fluid and crystallized intelligence (Jensen, 1998; Lynn & Vanhanen, 2002; Bouchard & Segal, 1985). Children from higher SES backgrounds tend to have higher amounts of *g*, fluid and crystallized intelligence, thought to reflect the high heritability of *g* (which itself affects SES), higher exposure to intelligence-enhancing experiences, and lower exposure to environmental factors that have a detrimental effects on brain development and/or cognitive test performance (Jensen, 1998, 2000, in Flynn, 2000; see Rapport & Weyandt, 2008 for a review of such factors). The effects of high SES and similar environmental factors were probably fairly homogeneous across the two samples, increasing the chances that a one-factor model would fit using CFA since a number of extraneous sources of variance were accounted for (i.e. approximately equated). The homogenizing effect of high SES in the current study's samples can be inferred from the relationships between Guess What and the other subtests. Guess What relies mainly on vocabulary and acquired knowledge with fewer

"reasoning" requirements (Reynolds & Kamphaus, 2009) and indeed, it had moderate to high correlations with the WISC-IV Vocabulary, Similarities, and Comprehension subtests in the clinical sample. Vocabulary, in turn, is highly positively correlated with SES (Heaton, Ryan, Grant, & Matthews, 1996; Sattler, 2001; Vernon, 1979). GWH had the highest correlations with all of the other RIAS and WISC-IV subtests, and was the only RIAS subtest to be significantly correlated with all of the WISC-IV subtests except Coding. Rather than indicating that all of these subtests involve verbal abilities or crystallized knowledge, these pervasive correlations may result from GWH being a proxy for homogeneously high SES, which is itself a proxy for a myriad of variables that impact the development of intellectual abilities. This argument is most defensibly applied to the typically-developing sample. However, its implications for interpreting the results of analyses of the clinical sample's data are less clear due to the heterogeneity of disorders and injuries, and the varying ages of onset and time since injury in this sample, all of which can impact the development of cognitive abilities in different ways at different rates (Dennis & Levin, 2004; Dennis, 2000) in ways that likely interact differently with genetic and environmental variables in idiosyncratic ways.

Invariance of the RIAS

The invariance of the one- and two-factor models was only assessed between the typically-developing group and a sub-sample of the clinical group with TBI, limiting the generalizability of conclusions drawn from these results about the construct validity of the RIAS to these groups. Nonetheless, the results added to concerns about the WHM subtest and provided evidence that comparison of latent means may not be supported between typically-developing children and those from at least one clinical group.

Specifically, the results revealed evidence of configural invariance for both models between the typically-developing group and the sub-group of individuals with TBI. This implies that, in each group, the number of factors and the same pattern of free and fixed parameters is similar across groups. The two-factor model was also invariant at the level of the factor loadings, providing evidence for metric invariance. This indicates that the verbal and nonverbal constructs have the same meaning for both groups; the strength of the relations between each subtest on the verbal and nonverbal factors is the same for both the TBI group and the typically-developing group (Byrne, 2010; Steinmetz, Schmidt, Tina-Booh, Wieczorek, & Schwartz, 2009).

On the other hand, the WHM factor loading had to be unconstrained in order to establish partial metric invariance. This suggests that WHM is not related to the *g* factor to the same extent in both groups. Indeed, its factor loading loaded highly on *g* in the TBI group but poorly on *g* in the typically-developing group. The higher loading of WHM in the TBI group may be a result of the effects of diffuse injury on multiple cognitive abilities, as discussed above. Since there are so few subtests on the RIAS, the lack of invariance of WHM between groups indicates that comparison of the CIX and/or NIX and/or WHM subtest may not be meaningful across TBI groups and typically-developing individuals. Calculating standard scores of individuals with TBIs based on normative data is effectively a cross-group comparison with a typically-developing sample. If the relationship between WHM and *g* is different for these two groups, it is unclear what the meaning of the WHM standard score is (and the NIX and CIX scores that are based on it).

Furthermore, the lack of scalar invariance (non-invariance of item intercepts) of both the one-factor and two-factor models across groups indicates that the measurement scales of the subtests have different intervals and zero points across groups. This means that latent means cannot be unambiguously compared across TBI and typically-developing groups because differences in the scales and origins of latent variables confound any effects of a between-group difference in latent means (Cheung & Rensvold, 2002). Differences in item intercepts are sometimes interpreted as systematic biases in the responses of groups to a subtest. As such, the "manifest mean can be systematically higher or lower (upward or downward biased) than one would expect due to the groups' latent mean and the factor loading" (Steinmetz et al., 2009, p. 603). All of the TBI groups' subtest intercepts were systematically lower than those of the typically-developing group. Although lower manifest means might be expected in a group of individuals with TBI (especially one that includes moderate and severe TBI), these lower intercepts indicate that these means may be lower than would be expected based on the factor loading and the TBI group's latent mean. However, it has been demonstrated that, even under factorial invariance, intercepts can still differ between groups as a result of actual group differences on a construct (e.g. Birnbaum, 1979; Millsap, 1998). It is predicted that the TBI group's factor mean scores on intelligence tests would be systematically lower than those of a typically-developing group, especially in the current study since the latter group appears to have intelligence scores higher than expected. Therefore, the lack of scalar invariance of the one- and two-factor models likely reflect actual group differences in the underlying factor rather than measurement bias of the RIAS.

Clinical Implications

The RIAS appears to measure crystallized, verbal abilities more clearly than fluid, spatial, nonverbal abilities. Thus, the validity of the NIX is questionable with concomitant problems with the validity of the CIX. The RIAS also seems to reliably produce higher scores than other intelligence measures, which may be particularly exaggerated in Canadian samples. These issues all have implications for the RIAS's application in clinical, educational, vocational, forensic, and organizational settings.

In the clinical sample (and in other studies), the consistently high *g*-loadings of the subtests provides some evidence of the CIX as an index of general intelligence. However, there were a substantial number of significant differences between the NIX and VIX scores in both clinical and typically-developing groups in the current study; it was almost normative for these differences to exist. When such significant differences exist, interpretation of the composite score is not always recommended, even though it is usually the more reliable score (Zhu & Weiss, 2005). However, interpreting the NIX and VIX separately is also problematic given the issues with the NIX's validity and with the nonverbal factor's low reliability (Beaujean et al., 2009). Furthermore, there were a greater proportion of individuals with significant differences between their nonverbal subtest scores than between their verbal subtest scores in the clinical and typically-developing groups, again calling into question how unitary is the construct that the nonverbal subtests are assessing. Specifically, significant divides between VRZ and GWH were found in 28.6% of the clinical sample and 34.7% of the typically-developing group compared with significant OIO/WHM differences in 59.8% of the clinical group and 54.5% of the typically-developing sample. With these differences, the advantage of

the CIX as a composite score more reliable than the NIX and VIX index scores becomes clear.

However, what is summarized by the CIX –that is, the abilities represented by it – may not be adequately representative of the intelligence domain. Instead, the subtests comprising the CIX may over-sample from the domain of verbal, crystallized intelligence. Does this matter when interpreting the CIX, as long as it can successfully predict outcomes? Gottfredson & Saklofske (2009) noted that:

> Tests that correlate more strongly with *g* are more construct valid for assessing general intelligence regardless of appearance, label, or intent. All major IQ tests measure *g* well, yet not equally well. Some yield verbally flavored full-scale IQs, others perhaps spatially flavored IQs. This imprecision in measuring *g* matters not for most practical purposes (e.g., selection, placement, vocational counselling), but greatly for others (e.g., exploring the brain correlates of *g*). The subtests of an IQ battery typically measure the general factor, *g*, to notably different degrees. (p. 190).

This suggests that it does not especially matter if some subtests are not as highly *g*-loaded or if the "nonverbal, fluid reasoning" subtests actually rely more heavily on verbal and crystallized abilities, as long as the FSIQ (or CIX) is useful in predicting outcomes. However, since the RIAS only has four subtests, it is of some concern that only two subtests, both of which are overtly verbal, load consistently highly on *g*. Tests with lower *g*-loadings have less predictive power (Jensen, 1998) and a "verbally-flavoured" CIX may place individuals who have stronger visual-spatial skills or more

impoverished environmental opportunities at a disadvantage compared to those with greater verbal abilities.

Indeed, it would seem that, as with any psychological test, the purpose for which an intelligence test's results will be used should dictate whether or not to use the RIAS, given its particular limitations. In general, it may be more prudent to select a longer intelligence test when high-stakes decisions (e.g. in certain forensic cases or when determining eligibility for services) are being based upon the results. On the other hand, the RIAS's advantages in time and cost savings over more traditional intelligence tests cannot be discounted and its psychometric properties are certainly strong enough to support its use for screening or research purposes, or to get a "general" idea of cognitive ability in clinical settings. However, there has been a greater movement in clinical work toward characterizing an individual's patterns of strengths and weaknesses, with more frequent use of domain-specific tests and examination of stratum II index over global IQ scores (e.g. VCI vs. PRI; Gottfredson & Saklofske, 2009). The RIAS's index scores would be unable to characterize diverse abilities if they are, in fact, comprised of subtests with items over-sampled from the crystallized intelligence domain.

Furthermore, choosing to use the RIAS instead of another, longer intelligence test (e.g. WISC-IV, Woodcock-Johnson-III-Cognitive) for diagnostic purposes will likely impact the diagnostic decision made in a substantial number of cases due to the RIAS' tendency to produce higher scores than other major intelligence tests.

Implications of Higher RIAS Index Scores

Although the RIAS and WISC-IV index scores were correlated, they were significantly different from each other, with the RIAS scores consistently higher than the

WISC-IV index scores, even after deleting outliers with large splits between these scores. Mean index differences between the WISC-IV and RIAS are not so surprising if the subtests that comprise the respective indexes are tapping different constructs. This latter point is especially apparent with the perceptual reasoning and nonverbal subtests. Even when taking the lowest value of the index difference confidence intervals, there is still a 5.19 standard point difference between the NIX and PRI and a 6.8 standard point difference between the CIX and FSIQ. This indicates that, in a clinical setting, even the lower limit of a confidence interval set around a CIX may be reliably higher than a mean FSIQ.

The DSM-IV-TR (2000) lists an IQ range of 50-55 to approximately 70 to meet diagnostic criteria for mild mental retardation. "Approximately" refers to the necessity of taking a standardized intelligence test's standard error of the mean into account which, for most tests, is about 5 standard score points above 70 (Schalock et al., 2010) to account for the standard error of measurement. This may mean that, when examining eligibility for intellectual disability, a given cut-off score (e.g. IQ = 70) is less likely to be included in the confidence interval for an individual administered a RIAS rather than the WISC-IV; this may impact access to services, even in the face of significant difficulties in adaptive functioning.

The finding that RIAS scores are higher than other intelligence test scores is consistent with previous studies that found higher RIAS index scores when compared with index scores on the Woodcock-Johnson-III-Cognitive in university students (Krach, Loe, Jones, & Farrally, 2009), on the WISC-IV in referred students, aged 6-12 years (Edwards & Paulin, 2007), on the WISC-III in referred students (McChristian, Windsor,

& Smith, 2007), in sub-samples of the normative group on the WISC-III and WAIS-III (Reynolds & Kamphaus, 2003) and on the WAIS-III in adults with intellectual disabilities (Umphress, 2008). In the latter study, 75% of the sample had CIX scores higher than FSIQ scores, 15% had equivalent CIX and FSIQ scores, and 10% had higher FSIQ than CIX scores. Furthermore, 35% of the sample would have been classified as having borderline intelligence based on their RIAS scores while their scores on the WAIS-III would have indicated mild intellectual disability. Even the General Ability Index (GAI), a score on the WISC-IV that de-emphasizes processing speed and working memory ability, was significantly lower than the CIX (Edwards & Paulin, 2007). This latter finding was in contrast to the RIAS's authors' (2003) contention that the RIAS is a more suitable measure of *g* since it relies less on psychomotor speed; when this variable is equated between the WISC-IV and RIAS, significant differences still remain. The authors of both of these studies called for caution in interpreting the NIX because of lower correlations with fluid, perceptual, and composite index scores, and recommended careful consideration when making placement and planning decisions based on the RIAS (Edwards & Paulin, 2007; Umphress, 2008).

Participant A, an outlier previously described, in the clinical group is a good illustration of the discrepancies that can result from choosing the RIAS versus the WISC-IV, for example. This participant's RIAS and WISC-IV scores were: VIX = 72, NIX = 41, CIX = 61; VCI = 45, PRI = 45, WMI = 62, PSI = 56, FSIQ = 42. The CIX would place this individual in the mild intellectual disability range while the FSIQ is in the moderate intellectual disability range. Furthermore, while the RIAS VIX indicates that this individual has a personal, relative strength in verbal abilities, the VCI indicates no

such strength in this area. Of course, more information would be gathered about this individual to complement the intellectual test's results, but the discrepancies between the RIAS and WISC-IV are illustrative of the differences test selection can make.

On the other hand, using the RIAS in the evaluation of individuals with suspected specific learning disabilities (LD) may actually make it easier to meet eligibility criteria when these criteria require a significant difference (e.g. 1.5 standard deviations) between intellectual functioning (i.e. an IQ score) and scores on a standardized test of achievement in the areas of reading, math, writing, etc. If the CIX is reliably higher than the FSIQ, the achievement score in a particular area would not have to be as low to be significantly discrepant. Although there has been a movement away from discrepancy-based definitions of learning disabilities (Flanagan, Ortiz, Alfonso, & Mascolo, 2002), these discrepancies might still be required to meet eligibility for services for some Canadian locales (Beal, 2004). In fact, even the Learning Disabilities Association of Canada (LDAC, 2002) has adopted a definition of learning disabilities which makes reference to "average abilities essential for thinking and/or reasoning" (LDAC, 2002). The definition also states that "learning disabilities are suggested by unexpected academic under-achievement or achievement which is maintained only by unusually high levels of effort and support" (LDAC, 2002). Both of these statements suggest that an LD is defined more easily in the presence of difficulties in achievement or domain-specific cognitive processing in the presence of average "thinking" ability (presumably as measured on an intelligence test). Thus the tendency of the RIAS to produce higher general intelligence scores than other intelligence tests may place more individuals in this "average" range, increasing the chances that they will meet this discrepancy criterion.

Meeting eligibility criteria for intellectual giftedness (often defined as having an IQ ≥ 130) would also be impacted by the choice of the RIAS over a Wechsler intelligence test. Since the intelligence test is sometimes the only criterion used to determine intellectual giftedness (McCoach, Kehle, Bray, & Siegle, 2001) more individuals given the RIAS would likely meet eligibility criteria for this designation than those who were administered a Wechsler intelligence test, for example. Although giftedness is thought to be characterized by advanced verbal abilities, problem-solving and abstract thinking abilities (Brueggemann, Reynolds, & Kamphaus, 2006), the RIAS's emphasis on crystallized abilities may penalize those who have not had as many opportunities to acquire knowledge while labelling as "gifted" those with superior verbal aptitude in the face of less than superior "raw" fluid reasoning.

The impact of test selection for meeting the eligibility criteria of intellectual giftedness can be illustrated with one of the outliers in the clinical group, Participant B. His scores on the RIAS and WISC-IV were as follows: VIX = 138, NIX = 122, CIX = 134; VCI = 128, PRI = 120, WMI = 100, PSI = 78, FSIQ = 112. This case may be illustrative of Reynolds and Kamphaus' (2003) contention that the RIAS is a measure of general intelligence with less reliance on psychomotor speed. Congruent with this individual's NIX and VIX scores, his/her VCI and PRI are in the superior range. However, his/her WMI and PSI are in the average and borderline ranges, respectively, which negatively impact his/her FSIQ, pulling it down to the average range. Calculated based on VCI and PRI only, a General Ability Index (GAI) of 127 is much more congruent with the CIX though the former would receive a classification of superior while the latter would place the individual in the very superior range. The difference

between the CIX and FSIQ in this individual is not trivial; the CIX could lead this individual to qualify for services under a "gifted" designation while interpretation of the FSIQ or even the GAI may not lead to such a classification. On the other hand, without subtests that measure psychomotor speed and working memory more directly, the RIAS was unable to capture this individual's weaknesses in these domains. If such testing were done in an educational setting (rather than in a more clinical setting), these domains may not be assessed directly on other tests. Expectations for the child could be very high based on his/her RIAS composite scores while not addressing or making accommodations for his/her potential limitations as found on the PSI and WMI scales of the WISC-IV.

The tendency of the RIAS to produce higher index scores than other intelligence tests may be especially exaggerated in Canadian samples when using American normative data. Indeed, when Canadian norms (unpublished) were created from the typically-developing group's data, an informal comparison revealed that the raw scores needed to achieve a *T*-score of 50 in this Canadian typically-developing sample corresponded to a substantially higher *T*-score when American norms were used. These differences were much more apparent in younger children (ages 4- 8 years) and on GWH and VRZ scores, suggesting that factors affecting crystallized intelligence (e.g. higher SES, earlier and/or higher-quality schooling, etc.; Wechsler, 2004) may account for the higher scores in Canadian children. Although Canadians often score significantly higher than Americans on intelligence tests (Wechsler, 1996, 2004), the score differences in the current sample's four to eight year olds ranged from an average of 8 to 16 standard points and so were higher than the 2.4 point difference found, for example, between the FSIQ of

American versus Canadian children on the WISC-IV (Wechsler, 2004). Thus, when using American normative data to calculate Canadians' scores, the RIAS index scores are likely to be inflated. This problem is magnified by the fact that the WISC-IV and WAIS-IV both have published Canadian normative data. If a clinician were to calculate an individual's FSIQ using Canadian normative data and the RIAS CIX using American norms, the difference between the scores is likely to be even greater than studies (including the current one) using only American norms have found. Thus, caution in selecting the RIAS for use with Canadians is especially warranted without Canadian normative data.

Limitations and Future Directions

The current study's findings must be interpreted with several caveats in mind, including: its use of only CFA and not EFA; the age range of participants; the variability in scores between samples; the relatively small sample sizes; the representativeness of the samples; the use of *T*-scores and standard scores in analyses; and the small number of indicators per factor in the confirmatory factor analyses.

Firstly, Frazier and Youngstrom (2007) have argued that CFA is being overly relied on in test development, leading to retention of an unmerited number of factors and, subsequently, the use of invalid index scores. Since only confirmatory factor analyses were completed, the current study did not allow for a direct comparison between the results of CFAs and EFAs on either data set. Certainly, the evidence for a one- or two-factor model would be more greatly supported if the results of both sets of analyses converged on the same model, something that has not happened in any previous studies, including those presented in the RIAS' technical manual (2003). However, it seems

unlikely that the results of CFA and EFA would produce the same factor solution for the RIAS; the data of a large group of referred students were analyzed by Beaujean, McGlaughlin, & Margulies (2009) with CFA, finding support for a two-factor model while Nelson, Canivez, Lindstrom, & Hatt (2007) used EFA on this same data set to provide evidence of a one-factor solution.

The current study only examined the factor structure of the RIAS in children and adolescents; it may be that the factor structure of the RIAS is different in adults and older adults. Dombrowski et al. (2009) found that, while the GWH and VRZ produced good *g*-loadings (≥.70) across the lifespan (3-94), the nonverbal subtests produced only fair *g*-loadings (.50 - .69) in the 6-18 year olds while in 19-94 year olds, good *g*-loadings of OIO and WHM were found. Furthermore, although the verbal subtests also loaded in the good range on the verbal subtest for most age groups, they loaded only in the fair range (.66-.69) in the 3-5 year olds. The variability in the factor loadings in EFA studies (Nelson et al., 2007; Dombrowski et al., 2009) is concerning. Since the current study included individuals across such a wide age range, it may be that the factor structure in the different age groups included (e.g. young children vs. adolescents) may differ, something which would be obscured in the current study and untestable with invariance testing given limited sample sizes in each age group. However, separate factor structures for different age groups were not postulated by the RIAS authors (Reynolds & Kamphaus, 2003), a contention supported by their finding of invariance of the RIAS' factor structure across multiple age groups in the normative sample. Nonetheless, caution should be taken when interpreting the results of the current study since children and adolescents across many developmental stages (from early childhood to late adolescence)

were included without the ability to test the RIAS' factor structure separately even for each developmental stage.

The clinical group included a very heterogeneous assortment of disorders and acquired injuries. Although it is important to establish the validity of the RIAS in a general clinical group, the heterogeneity made it difficult to make reference to underlying neurological and disease processes when attempting to hypothesize about the causes of the different model fits between this sample and the typically-developing sample. Ideally, future studies would be able to examine the RIAS's factor structure in a more homogeneous clinical group (e.g. traumatic brain injury; TBI) and control more for factors such as age at injury and time elapsed since injury to allow for integration between the intelligence and TBI literatures.

The necessity of performing the analyses on *T*-scores rather than raw scores due to the wide age range may have reduced variability in the data. It would be informative to obtain a large enough sample size of children of each age and analyze the factor structure of the RIAS in each age group using raw scores. However, the high factor loadings obtained using *T*-scores on the GWH and VRZ subtests is an indication that the factor structure of the test can still be examined, though the magnitude of the factor loadings may be affected.

Since there are a small number of indicators on the RIAS, there is a limit to the number and nature of models (e.g. hierarchical) that can be tested using CFA. This may be one advantage that EFA has in assessing the RIAS' factor structure, since it can be used to examine a hierarchical model, whereas such a model would be underidentified in the context of CFA due to the small number of indicators. Related to this, the fit indices

of the current study should be interpreted with the caveat that there are only one and two degrees of freedom separating the two-factor and one-factor model, respectively, from a fully saturated model (i.e. a model that accounts for all of the variance in the data). It is not entirely surprising that the model fit indices were so good simply because both tested models were so close to the saturated model. However, this cannot be remediated without adding subtests/indicators to the RIAS.

In the clinical group, there was a much greater range of scores than in the typically-developing group, especially in the lower end of the distribution. This likely increased the zero-order correlations between scores in the clinical group and may also have impacted the magnitude of factor loadings as well. However, the RIAS is used to assess individuals across a wide range of intellectual ability, so it is important that validation studies incorporate individuals with scores across this range.

It is unclear how representative participants in the typically-developing group were of Canadian children in general, or of children from Vancouver Island or Victoria. It is likely that these children were from a higher socioeconomic bracket than is found in Canada as a whole and the samples may not have been ethnically diverse enough to mirror Canadian demographics. For these reasons, caution is warranted when extrapolating from findings in these samples to Canadian children more generally.

Further studies are required to examine the validity of the RIAS's use in a clinical sample. Such studies should aim to elucidate the factor structure of the RIAS in more homogeneous clinical groups. Since EFA and CFA tend to produce disparate results, perhaps EFA could be used to explore the factor structure in one half of the sample while CFA is used to test the fit of models found from EFA methods in the second half of the

sample. If possible, future studies should try to assess the validity of the RIAS in smaller age groups, especially in young, developing samples.

RIAS scores were only compared with WISC-IV scores, a test normed for use with individuals aged 6 years to 16 years, 11 months. It may be that the pattern of correlations between scores would be different in younger children and/or in adults as assessed on an intelligence test that covers a broader age range, such as the Woodcock-Johnson-III Tests of Cognitive Abilities (WJ-III-Cog; Woodcock, McGrew, & Mather, 2001). Using the WJ-III-Cog is a cognitive test that was also explicitly modeled after the Cattell-Horn-Carroll theory of intelligence. If the RIAS subtests fit the factor structure of the WJ-III-Cog (with its G*f*, G*c*, and *g* factors), it would provide evidence for the validity of all of the RIAS' index scores.

The current study assessed the relationship between RIAS scores and scores on another intelligence test but did not examine the ability of the RIAS index scores to predict important academic and achievement outcomes. Future studies should aim to establish this kind of criterion-related validity, which is especially important in the case of the CIX since it has fewer subtests than other intelligence tests and two of these (OIO and WHM) appear to have lower and/or more variable *g*-loadings than might be desired given the established predictive power of *g* (Jensen, 1998; Edwards & Paulin, 2007). As well, the CIX's constituent tests de-emphasize processing speed and working memory, the latter of which is an especially reliable component of *g* (e.g. Colom, Rebollo, Palacios, Juan-Espinosa, & Kyllonen, 2004). For these reasons, the CIX may not be as reliable a measure of *g* as other intelligence tests. The RIAS's index and factor scores' ability to predict academic achievement in a group of referred and non-referred students

could be examined to establish the utility of the index scores in the face of equivocal evidence of their construct validity.

Summary

There have been several independent studies examining the validity of the RIAS since its publication. Due to differences in methodology (e.g. EFA versus CFA) and in characteristics of samples, the results of these studies have not always converged. Likewise, the current study produced equivocal findings and its results' generalizability is also limited by sample characteristics (e.g. heterogeneous clinical disorders, high SES Canadian sample drawn from a small geographic area, etc.). However, across all of these divergent studies, a pattern has emerged that has been echoed by the current study's findings. Namely, the validity of the NIX as a measure of nonverbal, fluid reasoning is questionable. The VIX appears to be a reliable, valid measure of verbal, crystallized abilities, and the CIX also shows evidence of being an index of general intelligence. Thus, while providing a good measure of G*c* and probably *g*, the RIAS's poor measurement of G*f* means that it does not line up well with the CHC model of intelligence. Therefore, the RIAS may be an adequate tool when practitioners need only a quick, broad indication of general intelligence in cases where important clinical decisions will not be based to a substantial degree upon its findings. Interpretation of VIX and NIX separately is inadvisable and it should be kept in mind that the CIX is likely "verbally-flavoured" and gives an advantage to those with higher amounts of crystallized knowledge while not necessarily capturing the abilities of those with more visual-spatial and/or fluid reasoning skills and/or with less environmental enrichment. Therefore, in cases where important diagnostic, forensic, and placement decisions will be

based in part on the results of an intelligence test, it may be advisable to select a test that has accrued better evidence that it assesses nonverbal reasoning abilities (G*f*) than the RIAS.

References

Alfonso, V. C., Flanagan, D. P., & Radwan, S. (2005). The impact of the Cattell-Horn-Carroll theory on test development and interpretation of cognitive and academic abilities. In D. P. Flanagan & P. L. Harrison (Eds)., *Contemporary Intellectual Assessment, Second Edition: Theories, Tests, and Issues* (pp. 185-202). New York, NY: Guilford Publications.

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

Arbuckle, J. L. (2010). AMOS (Version 19.0.0) [software]. Meadville, PA: SPSS.

Atkins v. Virginia, 260 Va. 375, 534S. E. 2d 312.

British Columbia Ministry of Education (2010). Special education services: A manual of policies, procedures and guidelines. Retrieved from BC Ministry of Education website: http://www.bced.gov.bc.ca/specialed/special_ed_policy_manual.pdf#page=47

British Columbia Statistics (2011).Profile of diversity in BC communities 2006. Retrieved from BCStats website: http://www.welcomebc.ca/local/wbc/docs/diversity/2006/Capital.pdf

Beal, A. L. (2004). Test review: Wechsler, D. (2003, 2004). Wechsler Intelligence Scale for Children Fourth Edition (WISC-IVCDN). Toronto, ON: The Psychological Corporation. *Canadian Journal of School Psychology, 19*(1/2), 221-234. doi: 10.1177/082957350401900112

Beaujean, A. A., McGlaughlin, S. M., & Margulies, A. S. (2009). Factorial validity of the Reynolds Intellectual Assessment Scales for referred students. *Psychology in the School, 46*(10), 932-950.

Bentler, P. M. (1990). Fit indices, LaGrange multipliers, constraint changes, and incomplete data in structural models. *Multivariate Behavioral Research, 25,* 163–172.

Bentler, P. M. (2004). *EQS: Structural equations program manual.* Encino, CA: Multivariate Software.

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

Binet, A. (1905). New methods for the diagnosis of the intellectual level of subnormals. *L'Annee Psychologique, 12,* 191-244. (Translated in 1916 by E. S. Kite in *The development of intelligence in children*. Vineland, NJ: Publications of the Training School at Vineland).

Binet, A. (1973). *Les idées moderne sur les enfants* [Modern ideas about children]. Paris: Flammarion (Original work published 1909).

Binet, A., & Simon, T. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'Année psychologique, 11*, 191-244.

Binet, A., & Simon T. (1909). L'intelligence des imbéciles. *L'Année Psychologique, 15*, 1-147.

Birnbaum, M. H. (1979). Procedures for the detection and correction of salary inequities. In T. R. Pezzullo & B. E. Brittingham (Eds.), *Salary equity* (pp. 121-144). Lexington, MA : Lexington Books.

Bock, G. R., Goode, J. A., & Webb, K. (2000). The nature of intelligence. *Novartis Foundation Symposium, 233*. Chichester: Wiley.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Bouchard, T. J., Jr., & Segal, N. L. (1985). Environment and IQ. In B. B. Wolman (Ed.), *Handbook of intelligence: Theories, measurements, and applications* (pp. 391-464). New York: Wiley.

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

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

Bulmer, Michael (2003). *Francis Galton: Pioneer of Heredity and Biometry*. Baltimore, MD: Johns Hopkins University Press.

Borkowski, J.G. (1985). Signs of intelligence: Strategy generalization and metacognition. In S. Yussen (Ed.), *The Growth of Reflection in Children* (pp. 105-144). Orlando, FL: Academic Press.

Brueggemann, A. E., Reynolds, C. R., & Kamphaus, R. W. (2006). The Reynolds Intellectual Assessment Scales (RIAS) and assessment of intellectual giftedness. *Gifted Education International, 21*, 127-136.

Burt, C. (1949). The structure of mind: A review of factor analysis. *British Journal of Educational Psychology, 19*, 100-111, 176-199.

Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Byrne, B. M. (2004). Testing for multigroup invariance using AMOS graphics: A road less traveled. *Structural Equation Modeling, 11*(2), 272-300.

Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (2nd ed.). New York, NY: Taylor & Francis Group.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105,* 456-466.

Campione, J. C., & Brown, A. L. (1978). Toward a theory of intelligence: Contributions from research with retarded children. *Intelligence, 2*, 279-304.

Carroll, J.B. (1982). The measurement of intelligence. In R.J. Sternberg (Ed.), *Handbook of human intelligence*. (pp. 3-28). Cambridge: Cambridge University Press.

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

Carson, J. (1993). Army alpha, army brass, and the search for army intelligence. *Isis, 84,* 278-309.

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

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

Ceci, S. J. (1996). *On Intelligence: A bio-ecological treatise on intellectual development*. 2nd ed. Cambridge, MA: Harvard University Press.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233-255. doi: 10.1207/S15328007SEM0902_5

Colom, R., Rebollo, I., Palacios, A., Juan-Espinosa, M., & Kyllonen, P. C. (2004). Working memory is (almost perfectly predicted by *g*. *Intelligence, 32*(3), 277-296.

Daniel, M. H. (1997). Intelligence testing: Status and trends. *American Psychologist, 52*(10), 1038−1045.

Daniel, M. H. (2000). Interpretation of intelligence test scores. In R. Sternberg (Ed.), Handbook *of intelligence* (pp. 477−491). New York: Cambridge.

Davenport, E. C., Jr. (1990). Significance testing of congruence coefficients: A good idea? *Educational and Psychological Measurement, 50*, 289-296.

Das, J. P., Naglieri, J. A., & Kirby, J. R. (1994). Assessment of cognitive processes: The PASS theory of intelligence. Boston, MA: Allyn & Bacon.

Dennis, M. (2000). Developmental plasticity in children: The role of biological risk, development, time, and reserve. *Journal of Communication Disorders, 33*(4), 321-332.

Dennis, M., & Levin, H. S. (2004). New perspectives on cognitive and behavioral outcome after childhood closed head injury. *Developmental Neuropsychology, 25*(1&2), 1-3.

Dennis, M., Francis, D. J., Cirino, P. T., Schachar, R., Barnes, M. A., & Fletcher, J. M. (2009). Why IQ is not a covariate in cognitive studies of neurodevelopmental disorders. *Journal of the International Neuropsychological Society, 15*(3), 331-343.

Detterman, D. K. (1987). Theoretical notions of intelligence and mental retardation. *American Journal of Mental Deficiency, 92* (1), 2-11.

Detterman, D. K. (1994). Toward an intelligent view of intelligence. *Psychological Inquiry, 5*(3), 201-203.

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

Dickens, W. T., & Flynn, J. R. (2001). Heritability estimates versus large environmental effects: The IQ paradox resolved. *Psychological Review, 108*(2), 346-369.

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development, 43*(2), 121-149.

Dombrowski, S. C. (2008). Review of RIAS. *Canadian Journal of School Psychology, 23*(2), 223-230.

Dombrowski, S. C., Watkins, M. W., & Brogan, M. J. (2009). An exploratory investigation of the factor structure of the Reynolds Intellectual Assessment Scales (RIAS). *Journal of Psychoeducational Assessment, 27*(6), 494-507.

Edwards, O. W., & Paulin, R. V. (2007). Referred students' performance on the Reynolds Intellectual Assessment Scales and the Wechsler Intelligence Scale for Children – Fourth Edition. *Journal of Psychoeducational Assessment, 25*(4), 334-340.

Flanagan, D. P., & Kaufman, A. S. (2004). Essentials of WISC-IV assessment. New York, NY: John Wiley and Sons Inc.

Flanagan, D. P., Ortize, S. O., Alfonso, V. C., & Mascolo, J. T. (2002). *The achievement test desk reference (ATDR)*. Boston: Allyn & Bacon.

Flynn, J. R. (2000). IQ gains, WISC subtests and fluid *g*: *g* theory and the relevance of Spearman's hypothesis to race. In G.R. Bock, J. A. Goode, K. Webb (Eds.) *The Nature of Intelligence*. Chichester, England: John Wiley & Sons Ltd, 202-216.

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

Galton, F. (1890). Kinship and correlation. *North American Review, 150*, 419-431.

Gardner, H. (1993). Frames of mind: The theory of multiple intelligences. New York, NY: BasicBooks.

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

Gottfredson, L., & Saklofske, D. H. (2009). Intelligence: Foundations and issues in assessment. *Canadian Psychology/Psychologie canadienne*, 50(3), 183-195. doi:10.1037/a0016641

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

Guttman, L. (1992). The irrelevance of factor analysis for the study of group differences. *Multivariate Behavioral Research, 27*, 175-204.

Heaton, R. K., Ryan, L., Grant, I., & Matthews, C. G. (1996). Demographic influences on neuropsychological test performance. In I. Grant and K. M. Adams (Eds.), *Neuropsychological assessment of neuropsychiatric disorders* (2nd ed.). New York, NY: Oxford University Press.

Hernstein, R., & Murray, C. (1994). *The bell curve*. New York, NY: Simon and Shuster.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179-185.

Horn, J. L. (1988). Thinking about human abilities. In J. R. Nesselroade (Ed.), *Handbook of multivariate psychology* (pp 645-685). New York: Academic Press.

Horn, J. L. (1991). Measurement of intellectual capabilities: A review of theory. In McGrew,

K.S., Werder, J.K., & Woodcock, R.W., *Woodcock-Johnson technical manual: A reference on theory and current research* (pp. 197-246). Allen, TX: DLM Teaching Resources.

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

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

Horn, J. L., & Stankov, L. (1982). Auditory and visual factors of intelligence. *Intelligence, 6*(2), 165-185.

Horn, J. L., Donaldson, G., & Engstrom, R. (1981). Apprehension, memory, and fluid intelligence decline in adulthood. *Research on Aging, 3*, 33-84.

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

Hunt, E. (1997). Nature vs. nurture: The feeling of vujà dé. In R. J. Sternberg & E. Grigorenko (Eds.), *Intelligence, heredity, and environment* (pp. 531-551). New York, NY: Cambridge University Press.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternate predictors of job performance. *Psychological Bulletin, 96,* 72–98.

Hunter, J. E., & Schmidt, F. L. (1996). Intelligence and job performance: Economic and social implications. *Psychology, Public Policy, and Law, 2,* 447–472.

IBM SPSS Statistics (Version 19.0.0) [software]. (2010). Somers, NY: SPSS Inc. an IBM company.

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

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36* (4), 409.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8 user's reference guide*. Chicago: Scientific Software International, Inc.

Kaplan, R. M., & Saccuzzo, D. P. (2005). *Psychological testing: Principles, applications, and issues, 6th edition*. Belmont, CA: Thomson Wadsworth.

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

Kaufman, A. S. & Lichtenberger, E. O. (2000). Essential of WISC-III and WPPSI-R assessment. New York, NY: John Wiley & Sons Inc.

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

Kelley, T. L. (1923). *Statistical method*. New York, NY: Macmillan.

Kenny, D. A. (2010). Structural equation modeling: Measuring model fit. Retrieved June 21, 2011 from http://www.davidakenny.net/cm/fit.htm

Krach, S. K., Loe, S. A., Jones, W. P., & Farrally, A. (2009). Convergent validity of the Reynolds Intellectual Assessment Scales (RIAS) using the Woodcock-Johnson Tests of Cognitive Ability, Third Edition (WJ-III) with university students. *Journal of Psychoeducational Assessment, 27*(5), 355-365.

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

Learning Disabilities Association of Canada (2002). *Official Definition of Learning Disabilities Adopted by the Learning Disabilities Association of Canada*, January 30, 2002. Retrieved from http://www.ldac-acta.ca/en/learn-more/ld-defined.html

Lezak, M. D., Howieson, D. B., & Loring, D. W. (2004). The Rationale of deficit measurement. In M. D. Lezak, D. B. Howieson, & D. W. Loring (Eds.), *Neuropsychological assessment* (4th ed.) (pp. 86-99). New York, NY: Oxford University Press.

Loehlin, J. C. (1998). *Latent variable models: An introduction to factor, path, and structural analysis* (3rd ed.). Mahwah, NJ: Erlbaum.

Lynn, R., & Vanhanen, T. (2002). *IQ and the wealth of nations*. Westport, CT: Praeger.

McCoach, D. B., Kehle, T. J., Bray, M. A., & Siegle, D. (2001). Best practice in the identification of gifted students with learning disabilities. *Psychology in the Schools, 38*, 403-411.

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.

McGrew, K. S. (2005). The Cattell–Horn–Carroll theory of cognitive abilities. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 136−181), 2nd ed. New York: Guilford Press.

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

McGrew, K. S., Flanagan, D. P., Keith, T. Z., & Vanderwood, M. (1997). Beyond *g*: The impact of Gf-Gc specific cognitive abilities research on the future use and interpretation of intelligence tests in the schools. *School Psychology Review, 26,* 177– 189.

McGrew, K. S., Werder, J. K., & Woodcock, R. W. (1991). *The WJ-R technical manual*. Chicago: Riverside.

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.

Millsap, R. E. (1998). Group differences in regression intercepts: Implications for factorial invariance. *Multivariate Behavioral Research, 33*(3), 403-424.

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., et al. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*, 77–101.

Nelson, J. M., Canivez, G. L., Lindstrom, W., & Hatt, C. V. (2007). Higher-order exploratory factor analysis of the Reynolds Intellectual Assessment Scales with a referred sample. *Journal of School Psychology, 45*, 439-456.

Rapport, L. J., & Weyandt, L. L. (2008) The development and measurement of intelligence. In J. M. Sattler (Ed.), *Assessment of children: Cognitive foundations* (pp. 245-264). La Mesa, CA: Author.

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

Reynolds, C. R., & Kamphaus, R. W. (2009). Development and application of the Reynolds Intellectual Assessment Scales (RIAS). In J. A. Naglieri, & S. Goldstein

(Eds.), *Practioner's guide to assessing intelligence and achievement.* Hoboken, NJ: John Wiley & Sons, Inc.

Rimoldi, H. J. A. (1948). Study of some factors related to intelligence. *Psychometrika, 13*(1), 27-46.

Sattler, J. M. (2001). *Assessment of children: Behavioral and clinical applications* (4th ed.). La Mesa, CA: Author.

Sattler, J. M. (2008) Assessment of children: Cognitive foundations. La Mesa, CA: Author.

Sattler, J. M., & Dumont, R. (2004). Assessment of children: WISC-IV and WPPSI-III supplement. Le Mesa, CA: Author.

Schalock, R. L., Borthwick-Duffy, S. A., Bradley, V. J., Buntinx, W. H. E., Coulter, D. L., Craig, E. M.…& Yeager, M. H. (2010). *Intellectual disability: Definition, classification, and systems of supports* (11th ed.). Washington, DC: American Association on Intellectual and Developmental Disabilities.

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

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*(2), 262-274.

Schneider, J. (July 1, 2011). *Is g and ability?* Retrieved from http://assessingpsyche.wordpress.com/page/2/

Schraw, G. (2005). Review of the Reynolds Intellectual Assessment Scales and the Reynolds Intellectual Screening Test. In R. A. Spies, & B. S. Plake (Eds.), *The*

sixteenth mental measurements yearbook (pp. 894−895). Lincoln, NE: Buros Institute of Mental Measurements.

Shaywitz, B. A., Holford, T. R., Holahan, J. M., Fletcher, J. M., Stuebing, K. K., Francis, D. J., & Shaywitz, S. E. (1995). A Matthew effect for IQ but not for reading: Results from a longitudinal study. *Reading Research Quarterly, 30*(4), 894-906.

Siegler, R. S. (1994) Cognitive variability: A key to understanding cognitive development. *Current Directions in Psychological Science, 3*(1), 1-5. Retrieved from http://www.jstor.org/stable/20182248

Snow, R. E. (1998). Abilities and aptitudes and achievements in learning situations. In J. J. McArdle, & R. W. Woodcock (Eds.), *Human cognitive abilities in theory and practice*

(pp. 93−112). Mahwaw, NJ: Lawrence Erlbaum.

Spearman, C. (1904). "General intelligence," objectively determined and measured. *The American Journal of Psychology*, 201-292.

Spearman, C. (1923, 1973). *The nature of intelligence and the principles of cognition*. Arno Press, New York .

Stanovich, K.E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*, 360 – 407.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25*,173–180.

Steinmetz, H., Schmidt, P., Tina-Booh, A., Wieczorak, S, & Schwartz, S. H. (2009). Testing measurement invariance using multigroup CFA: Differences between

educational groups in human values measurement. *Quality and Quantity, 43*, 599-616. doi: 10.1007/211135-007-9143-x

Sternberg, R. J. (1985). Beyond IQ: A triarchic theory of human intelligence. New York, NY: Cambridge University Press.

Sternberg, R. J., & Kaufman, J. C. (1998). Human abilities. *Annual Review of Psychology, 49*, 1134−1139.

Sugawara, H. M., & MacCallum, R. C. (1993) Effect of estimation method on incremental fit indexes for covariance structure models. *Applied Psychological Measurement, 17*, 365-377.

Tabachnick, B.G. and Fidell, L.S. (2007). Using Multivariate Statistics (5th ed.). New York: Allyn and Bacon.

Thomson, G. H. (1951). *The factorial analysis of human ability* (5th ed.). London: University of London Press.

Thurstone, L. L. (1904). *Introduction to the theory of mental and social measurements.* New York: Columbia University, Teachers College.

Thurstone, L. L. (1931). *The reliability and validity of tests*. Ann Arbor, MI: Edwards Brothers.

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

Umphress, T. (2008). A comparison of low IQ scores from the Reynolds Intellectual Assessment Scales and the Wechsler Adult Intelligence Scale – Third Edition. *Intellectual and Developmental Disabilities, 46*(3), 229-233.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41*, 321-327.

Vernon, P. E. (1979). *Intelligence: Heredity and environment.* San Francisco, CA: Freeman.

Wechsler, D. (1996). *Wechsler Intelligence Scale for Children- Third Edition Canadian technical manual*. Toronto: The Psychological Corporation.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children – Fourth Edition. technical and interpretive manual*. San Antonio, TX: Harcourt Assessment, Inc.

Wechsler, D. (2004). *Wechsler Intelligence Scale for Children–Fourth Edition: Canadian Manual.* Toronto, Ontario, Canada: PsychCorp.

Willoughby, R. R. (1927). Family similarities in mental-test abicalities (with a note on the growth and decline of these abilities). *Genetic Psychology Monographs, 2,* 235-277.

Wissler, C. (1901). The correlation of mental and physical traits. *Psychological Monographs, 3*, 1-62.

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

Yerkes, R. M. (Ed.) (1921) Psychological examining in the United States Army. *Memoirs of the National Academy of Sciences, 15,* 1-890.

Yoakum, C., & Yerkes, R.M. (1920). *Army mental tests*. New York: Henry Holt and Company.

Zhu, J., & Weiss, L. (2005). The Wechsler scales. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests and issues* (pp. 297-324). New York, NY: The Guilford Press.