Running head: MONTE CARLO MODELING OF IQ TEST STRUCTURE

This document is consistent with, but may not exactly match, the version of scholarly record published online by Sage Publications in *Assessment* on August 20, 2019 at the following link: *https://doi.org/10.1177/1073191119869828*

Monte Carlo Modeling of Contemporary Intelligence Test (IQ) Factor Structure: Implications for IQ Assessment, Interpretation and Theory

Stefan C. Dombrowski
Rider University
sdombrowski@rider.edu

Ryan J. McGill
College of William & Mary
rmcgill@wm.edu

Grant B. Morgan
Baylor University
Grant_Morgan@baylor.edu

Declarations of interest: None.

Correspondence: Stefan C. Dombrowski, Ph.D., Professor & Director, School Psychology Program, Department of Graduate Education, Leadership & Counseling. Rider University, 2083 Lawrenceville Road, Lawrenceville, NJ 08648. sdombrowski@rider.edu

Abstract

Researchers continue to debate the constructs measured by commercial ability tests. Factor analytic investigations of these measures have been used to develop and refine widely adopted psychometric theories of intelligence particularly the Cattell-Horn-Carroll (CHC) model. Even so, this linkage may be problematic as many of these investigations examine a particular instrument in isolation and CHC model specification across tests and research teams has not been consistent. To address these concerns, the present study used Monte Carlo (MC) resampling to investigate the latent structure of four of the most widely used intelligence tests for children and adolescents. The results located the approximate existence of the publisher posited CHC theoretical group factors in the DAS-II and the KABC-II but not in the WISC-V or the WJ IV Cognitive. Instead, the results supported alternative conceptualizations from independent factor analytic research. Additionally, whereas a bifactor model produced superior fit indices in two instruments (WISC-V and WJ IV Cognitive), a higher-order structure was found to be superior in the KABC-II and the DAS-II. Regardless of the model employed, the general factor captured a significant portion of each instrument's variance. Implications for IQ test assessment, interpretation and theory are discussed.

Keywords: Monte Carlo Simulation; Modeling; IQ tests; Intelligence; Factor Structure; Cattell-Horn-Carroll; Factor Analysis

Monte Carlo Modeling of Contemporary Intelligence Test (IQ) Factor Structure: Implications for IQ Assessment, Interpretation and Theory

The attempt to ascribe a well-substantiated theory and interpretive approach that can be applied consistently across intelligence (IQ) tests is laudable but remains elusive. The construct of IQ is reflected by a latent, unobservable variable and/or set of variables and researchers have been debating how to “best” measure the construct for more than a century (Carroll, 1993). Despite often polemic debates over the past two decades about which model and interpretive approach is superior, researchers forge ahead and try to fit theoretical and applied models to the construct of intelligence; sometimes researchers become entrenched in dogmatic, polarizing positions about which theoretical conceptualization, model, and interpretive approach is “best” (Benson, Beaujean, McGill & Dombrowski, 2018; McGill, Dombrowski & Canivez, 2018).

One of the more empirical ways to understand the nature of intelligence is to investigate the theoretical structure presumed to undergird our tests of cognitive ability. Typically, this is accomplished through factor analysis. Although it is recognized that intelligence should not solely be defined in a narrow, mechanistic way, as this will overlook other dimensions of intellectual ability, the use of factor analysis portends to offer insight into understanding the construct of intelligence, the theories related to it, and the interpretive methods used to understand the constructs measured by tests thought to reflect that psychological attribute.

Publishers of commercial ability measures present detailed information within their technical manuals including an overview of an instrument’s theoretical orientation and validity analyses. One type of validity, structural or factorial validity, is given extensive coverage in a test’s technical manual and may be established through the use of exploratory and/or confirmatory factor analyses. Structural validity is arguably the most important type of validity evidence as it sets the stage for all other aspects of validity (Keith & Kranzler, 1999). Put simply, structural validity determines whether an instrument is consistent with theory and subsequently how an instrument should be interpreted.

Millions of individuals are administered IQ tests annually. Many of these administrations involve high stakes decisions for children in a K-12 setting; thus, the importance of having a stable test structure is of paramount importance. Over the past two decades independent research has questioned the theoretical structure of frequently administered IQ tests (Canivez, 2008; Canivez, Watkins & Dombrowski, 2018; Dombrowski, 2013; Dombrowski, Watkins & Brogan, 2009; Dombrowski, McGill & Canivez, 2017, 2018) and a significant amount of research attention has focused particularly on cognitive measures that may be administered to children and adolescents. The majority of these investigations have yielded a structure divergent not only with that found in the test publisher’s technical manual, but also with other independent studies (e.g., Canivez & Watkins, 2016; Dombrowski, McGill & Canivez, 2018; Dombrowski, McGill, Canivez & Peterson, 2019; Keith, Low, Reynolds, Patel & Ridley, 2010; McGill & Dombrowski, 2018; Reynolds & Keith, 2017; Reynolds, Keith, Fine Fisher & Low, 2007). This is cause for considerable concern, and indicates a possible replication problem (Cronbach & Meehl, 1955). If a test does not measure what it purports to measure, or measures a construct differently depending upon the researcher investigating the instrument or the model/method used by the researcher, it may be difficult to have confidence in the instrument’s underlying theoretical structure and in the applied, interpretive conclusions that may be drawn from it (Gorsuch, 1983; Meehl, 1990).

Review of the Structural Validity Literature

As previously mentioned, there is a considerable body of independent factorial validity research from the past two decades that has questioned the theoretical structure of many commercial IQ tests. This article will primarily review the structural validity evidence from four of the most frequently administered IQ tests for children and adolescents (Sotelo-Dynega & Dixon, 2014) since they will be the subject of inquiry in this article: the Wechsler Intelligence Scale for Children-Fifth Edition (WISC-V; Wechsler, 2014), Woodcock-Johnson IV Tests of Cognitive Abilities (WJ IV Cognitive; Schrank, McGrew & Mather, 2014), Kaufman Assessment Battery for Children-Second Edition (KABC-II; Kaufman & Kaufman, 2004), and the Differential Abilities Scales-Second Edition (DAS-II; Elliot, 2007). For the sake of parsimony, we will focus our review on factor analytic studies that have used the total test battery (core and supplemental subtests) as the basis for investigation since this more completely permits an analysis of linkage with CHC theory.

WISC-V. Canivez and Watkins (2016) used both exploratory factor analysis (EFA; the Schmid-Leiman [SL; Schmid & Leiman, 1957] approximate bifactor analysis) and confirmatory factor analysis (CFA) to investigate the total WISC-V battery. Canivez and Watkin's study suggested that the WISC-V is a four factor instrument reflecting one general and four group factors (e.g., Verbal Comprehension (Gc), Working Memory (Gsm), Processing Speed (Gs), and Perceptual Reasoning [PR]). It should be noted that this model is consistent with previous Wechsler theory. Canivez and Watkins first used EFA to guide the determination of the instrument's factor structure and relied upon subsequent CFA analyses to establish the best model fit. Canivez and Watkins compared numerous competing models including all of the models presented in the technical manual. Their results suggested that a four bifactor model, consistent with their EFA results, had the best model fit. This same result was replicated by additional researchers who used both the normative data set and independent clinical samples (Canivez, McGill, Dombrowski, Watkins, Pritchard & Jacobson, 2018; Canivez, Watkins & Dombrowski, 2016; Dombrowski, Canivez & Watkins, 2018).

Even so, there are two WISC-V studies that found a different structure from that of Canivez and Watkins as well as the test publisher. Dombrowski, Canivez, Watkins and Beaujean (2015) used exploratory bifactor analysis (EBFA; Jennrich & Bentler, 2011) and located three group factors and one general factor. Curiously, the Verbal Comprehension factor was not located. Dombrowski et al. (2015) opined that the reason for this anomalous result was that the verbal subtests, which themselves are predominantly *g* loaded, collapsed onto the general factor. They acknowledged that this finding was inconsistent with decades of research into the Wechsler scales and could be an artifact of the recently created EBFA method that tends to display group factor collapse in the presence of a strong general factor (Mansolf & Reise, 2016).

In contrast, Reynolds and Keith (2017) used CFA to argue that the WISC-V's structure was hierarchical, containing five group factors plus a general factor and was essentially consistent with the theoretically proposed structure (e.g., Gc, Gf, Gv, Gsm & Gs)[1] presented in

---

[1] Throughout this article the following may serve as a legend for the CHC abbreviations: Crystalized Intelligence/Knowledge [Gc], Fluid Intelligence/Reasoning [Gf], Visual-Processing [Gv], Short-Term Memory/Working Memory [Gsm], Long-Term Retrieval [Glr], Processing Speed [Gs], and Auditory Processing [Ga]

the technical manual. Reynolds and Keith engaged in a series of adjustments (~20) to arrive at their final, validated model. Specifically, this entailed having the Arithmetic subtest load on both the Working Memory factor as well as the second-order general factor. Second, Reynolds and Keith (2017) also correlated the disturbance between the Gv and Gf group factors[2]. Whereas some researchers support circumspect model post hoc model fitting (e.g., Bryne, 2005), others caution against its use (e.g., Brown, 2015; Cucina & Howardson, 2017; Kerr, 1998; Kline, 2016). To be fair, there is nothing inherently wrong with this practice; however, it should be noted that *exploratory* model fitting attempts may produce structures that capitalize on chance and may not replicate under more stringent measurement conditions (MacCallum, Roznowski, & Neocowitz, 1992; Meehl, 1978). In a subsequent CFA analysis of the Canadian version of the WISC-V by Watkins, Dombrowski, and Canivez (2018), the higher-order model posited by Reynolds and Keith (2017) did not yield the best fit to the data. Instead results supported a four-factor bifactor model consistent with Canivez and Watkins (2016).

WJ IV Cognitive. The only extant research to independently investigate the WJ IV Cognitive's structure was that reported by Dombrowski, McGill and Canivez (2017, 2018). Dombrowski et al. conducted two separate studies: an EFA study using the SL approximate bifactor analysis followed by a CFA study to investigate the results determined by EFA. Across both studies, these researchers contended that a one general and four group (e.g., Gc, PR, Gsm, and Gs) factor approximate SL bifactor structure, quite different from the test publisher's proposed seven factor higher order structure (i.e., Gc, Gf, Gv, Gsm, Glr, Ga, and Gs), was the structure with the best global and local fit at ages 9-13. The structure uncovered by Dombrowski et al. was reminiscent of the prior theoretical structure for the Wechsler Scales (e.g., WISC-IV; Wechsler, 2003). Thus, Dombrowski et al. questioned not only the theoretical orientation of the WJ IV Cognitive but also whether the instrument should be interpreted the way the technical manual suggested it should be interpreted.

KABC-II. Reynolds, Keith, Fine, Fisher and Low (2007) used a higher order CFA to investigate the structure of the KABC-II. Reynolds et al. incorporated only three adjustments to improve upon the test publisher's model. This included correlating the error terms between Atlantis and Atlantis Delayed and Rebus and Rebus Delayed; specifying secondary loadings for the Hand Movements and Pattern Reasoning subtests; and placing Gestalt Closure on the Gv rather than the Gc factor, which is inconsistent with publisher theory. Thus, Reynolds et al. corroborated the existence of the test publisher's Cattell-Horn-Carroll (CHC; Carroll, 1993; Horn & Cattell, 1966; Schneider & McGrew, 2018) five factor higher-order theoretical structure (e.g., Gc, Gf, Gv, Glr, and Gsm). Desired simple structure was attained with the few noted exceptions. On the other hand, McGill and Dombrowski (2018) using the SL approximate bifactor procedure found that the structure was significantly different from that proposed in the Technical Manual. McGill and Dombrowski's exploratory study revealed one general factor and four group factors (e.g., Gc, PR, Gsm, and Glr) structure. Whereas Reynold's et al.'s study obtained a structure largely consistent with that theoretically proposed in the technical manual, McGill and

[2] In previous WISC-V CFA studies, including those furnished in the Technical Manual, The path between *g* and Gf often equaled or exceeded unity suggesting an impermissible solution. Addition of this parameter appeared to resolve that issue.

Dombrowski's study yielded a structure that deviated quite significantly from publisher theory. Nonetheless, simple structure was readily attained in McGill and Dombrowski's model.

DAS-II. Both Keith et al. (2010) and Dombrowski, McGill, Canivez and Peterson (2018) independently investigated the structure of the DAS-II. Dombrowski et al used the SL approximate bifactor procedure and located six distinct CHC group factors (Gc, Gf, Gv, Gsm, Gs, and Glr); however, Early Number Concepts (Gf/Gc) and Picture Similarities (Gf) saliently loaded only on the general factor. All other subtests were consistent with their theoretically posited factors and Dombrowski et al. concluded that simple structure was attained. Earlier, Keith et al. investigated the structure of the DAS-II using higher order confirmatory factor analysis. The results of Keith et al.'s study suggested that the DAS-II reflects six higher order factors consistent with that proposed by the publisher. Keith et al. incorporated several adjustments to arrive at the structure suggested by the test publisher. This included the need to correlate the disturbance terms between Gf and Gv, and place Verbal Comprehension on both the Gc and the Gf factors. Additionally, Keith et al. incorporated a correlated disturbance term between Recall of Designs and Copying. Although Keith et al.'s study located the six posited CHC factors, it may be viewed as less parsimonious than that proposed by Dombrowski et al because of need for the additional model adjustments. Still, both studies located six CHC group factors with varying degree of consistency.

Summary. Some studies have offered dramatically different results from that proposed within the instruments' technical manuals; other studies have offered results demonstrating a greater consistency with posited CHC structures. Nevertheless, in almost all cases, the final structure found within the independent research has diverged from that proposed by the test publisher and, in some cases, even that produced by other independent research. Complex parameters are specified for certain tests (e.g., Reynolds & Keith, 2017) that are not consistently modeled or explored on other tests that purport to measure the same latent constructs. Given these discrepancies, a factor analytic investigation, using multiple ability measures, and incorporating *all* posited models for each of those measures in the form of a psychometric meta-analysis (i.e., Carroll, 1993) would be instructive for determining the nature and structuring of CHC variables across tests and, more generally whether a meta-theory largely derived from research summaries of the factor analytic literature such as CHC is viable (Meehl, 1990). This will reflect the first time that an investigation of this nature will be furnished in the literature.

A Replication Crisis and a Possible Way Forward?

In none of the above final, validated models do we see an exact replication of the CHC theoretical structure proposed by the publisher nor do we see independent research conclusions completely agree with each other. Disparate results from the same datasets and even the same theory suggest the potential of a replication crisis (Cronbach & Meehl, 1955; Meehl, 1978) as seen in other areas of psychological science and other disciplines (Shrout & Rodgers, 2018).

What conclusion from this disparate literature are we to trust for a particular instrument? Which study offers a way forward for the field? Certainly, both global and local fit are important arbiters along with parsimony and theory. Gorsuch (1983) noted that when multiple methods of factor analysis converge upon a structure then we can be *more* confident that it reflects the true structure for a given covariance matrix.

Although no method of modeling or no specific model can be viewed as beyond reproach (Box & Draper, 1987; Kline, 2016), Monte Carlo simulation can offer an elegant way forward by permitting an analysis of the structure of an instrument over numerous replications (Casey & Harden, 2014). Simulation permits investigation of not just one study at a time but rather hundreds, if not thousands, limited only by computing power. As a result it may well be a powerful tool to compare the theoretical structure of major IQ tests among the competing models offered by independent research and the test publisher, overcoming sampling limitations that frequently encumber factor analytic investigations.

Accordingly, the purpose of this study is to use Monte Carlo modeling to investigate the theoretical structure of the WISC-V, WJ IV Cognitive, KABC-2 and DAS-II, all of which were all developed to align with CHC theory. As consensus on what these instruments measure remains elusive, simulation may be useful for comparing the results furnished across the psychometric literature and shedding insight into whether posited models are likely to be replicated across different measurement conditions.

Method

Data analyses involved several steps. First, the normative sample correlation matrices from WISC-V, DAS-II, KABC-II and WJ IV Cognitive were obtained from available technical manuals and/or extracted from normative datasets with permission from the test publisher. For the WISC-V, the correlation matrix for the total sample across the 6 to 16 age range was available in the instrument's technical manual. Likewise, the WJ IV Cognitive technical manual provided the intercorrelation matrix among all subtests across all age ranges. The matrices for ages 9 to 13 were utilized in this study as they were identical to the ones used by Dombrowski et al. (2017, 2018). With the DAS-II and KABC-II, the respective technical manuals did not provide an intercorrelation matrix for participants who were administered every subtest. However, participants in the standardization sample across the 5 to 8 and 7 to 18 age ranges, respectively, were administered every core and supplementary subtest in the instrument. Thus, the standardization data was used to create the correlation matrices for these measures at those age brackets. The resulting correlation matrices from the four instruments were used as the basis for MC simulation.

Data Analyses

Monte Carlo Simulation. The next step was to simulate the correlation matrix of each of the measures 1,000 times using Monte Carlo (MC) resampling methods (Metropolis & Ulam, 1949). Resampling involves drawing multiple random samples of data from an assumed data generating process (DGP). In comparison to other simulation techniques, MC resampling draws samples in an iterative process from the observed data rather than a researcher-derived DGP. According to Casey and Harden (2014), if the observations in the observed data represent a random sample from the larger population, the samples generated through MC simulation can be assumed to approximate the distribution from this broader population. The sample size used for simulation was that reported in either the various instrument's technical manuals or standardization sample: WISC-V (age 6-16, N = 2,200); WJ IV Cognitive (age 9-13, n = 1,572); DAS-2 (age 5-8, n = 787); and KABC-II (age 7-18, n = 1,142).

Confirmatory Factor Analyses. Next, best fitting CHC measurement models that have been reported in the professional literature were tested for each instrument using CFA to determine which theoretical structure provided the best fit across 1,000 replications. For the

WISC-V, publisher theory, Canivez and Watkin's (2016) four factor bifactor structure, and Reynolds and Keith's (2017) five factor higher-order structure were examined. Also evaluated was Dombrowski et al.'s (2015) exploratory bifactor factor structure (one general and three group factors). For the WJ IV Cognitive, the BF models produced by Dombrowski and colleagues (2017, 2018) were examined along with the seven-factor higher-order model posited by the WJ IV Cognitive test publisher. For the DAS-II, publisher theory was explicated in addition to Keith et al.'s (2010) six factor higher-order structure alongside Dombrowski et al.'s (2018) EFA/SL approximate BF five factor structure across ages 5-8[3]. Finally, for the KABC-II, a five factor SL approximate bifactor structure from McGill and Dombrowski (2018) and Reynolds et al. (2007) five factor higher order structure was evaluated across the age 5 to 8 time span in comparison to publisher theory.

Model Fit. Global fit was examined iteratively using the models posited by publishers as a baseline. The χ^2 value was one of several metrics used for model comparison. Lower values generally indicate better model fit. Because of concerns about distortion of the χ^2 value when larger samples are used (e.g., Kline, 2016), approximate fit indices were also referenced to further evaluate model selection (see Lai & Green, 2016). The root mean square error of approximation (RMSEA), comparative fit index (CFI), and Tucker-Lewis index (TLI) were used to comprehensively assess all aspects of global model fit. Although universally accepted criterion values for approximate fit indices do not exist (McDonald, 2010), generally accepted guidelines were referenced (Hu and Bentler, 1999). In general, higher values indicate better fit for the CFI and TLI while lower values suggest better fit for the RMSEA. In particular, Hu and Bentler's (1999) decision-making rules were considered: CFI and TLI $\geq .90$ combined with RMSEA $\leq .08$ were criteria for adequate model fit while CFI/TLI ≥ 0.95 and RMSEA ≤ 0.06 were indicative of good model fit. The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were also considered. Because AIC and BIC do not have a meaningful scale, the model with the smallest AIC and BIC values were preferred, and considered the model most likely to replicate (Kline, 2016). Superior models required adequate to good overall fit and the indication of meaningfully better fit (ΔCFI > .01, ΔRMSEA > .015, ΔAIC and ΔBIC >10) than alternative models (Burnham & Anderson, 2004; Cheung & Rensvold, 2002; Chen, 2007). Local fit was also considered, particularly parsimony of structure and utility of the model as basis for the creation of interpretive indices, as models should never be retained or rejected exclusively on the basis of global fit testing (Brown, 2015; Byrne, 2005).

Metrics of interpretability were estimated and evaluated including coefficients omega-hierarchical (ω_H) and omega-hierarchical subscale (ω_{HS}). Although ω_H and ω_{HS} were initially conceptualized as model-based reliabilities and estimates of the reliability of unit-weighted scores produced by a given set of indicators (Reise, 2012; Rodriguez, Reise & Haviland, 2016; Watkins, 2017), they may also be conceptualized as a metric of construct interpretability of the resulting latent factors (Gustafsson & Åberg-Bengtsson, 2010). For most conventional IQ tests,

[3] With the exception of the WISC-V, which included all normative participants, analyses of other measures were restricted to certain age brackets in order to comport with results reported in the professional literature. For example, the DAS-II and KABC-II contain multiple subtest measures that can only be administered at specific ages (i.e., early childhood). As a result, Keith et al. (2010) used the data from 5-8 age bracket as a focal point of their analyses on the DAS-II because those were the only participants who were administered all DAS-II subtests thus maximizing potential linkages to CHC theory.

the ω_H coefficient represents the estimate of the general intelligence factor variance with variability from the group factors removed; the ω_{HS} coefficient represents the group factor estimate with variability from all other group *and* general factors removed (Brunner, Nagy, & Wilhelm, 2012; Reise, 2012). In general, the ω_H/ω_{HS} coefficient thresholds needed for confident clinical interpretation of general and group-factor based scales is a value of .50 but .75 is generally preferred (Reise, 2012; Reise et al., 2013). Additionally, Hancock and Mueller's (2001) coefficient H was also calculated. The *H* coefficient provides an estimate of how well a posited latent construct is (a) represented by a given set of indicators and (b) likely to be replicated across similar factor analytic studies. A value of .70 or higher is preferred for interpretation of factor indices (Hancock & Mueller, 2001; Rodriguez et al., 2016). Finally, the percentage of uncontaminated correlations (PUC) was estimated to provide a sense of how many of the correlations in the model inform directly on the general factor. A PUC value of .70 or higher (along with explained common variance of .70 or higher) would suggest the measure is primarily unidimensional.

Mplus 8.2 (Muthén and Muthén, 2017) was used first to simulate 1,000 datasets (i.e., replications) by essentially sampling from the population to which the normative correlation matrix applies for each of the aforementioned instruments and then to apply CFA using maximum likelihood (ML) estimation. The number of replications should be chosen based on the purpose of the study, reach desired level of sampling variance, and/or obtain adequate power (Bandalos & Leite, 2013; Harwell, Stone, Hsu, & Kirisci, 1996). The reported range of replications used in studies involving latent variable models, such as this one, are 20 to 1,000, with a median of 200 (Powell & Schafer, 2001). By using 1,000 replications for each normative correlation matrix in the current study, all average correlations across the 1,000 sampled correlation matrices matched the normative correlation matrix to at least three decimal places for each of the instruments examined. Omega coefficients, *H*, and PUC were calculated based upon the average of the standardized factor loadings across the 1,000 replications using Watkins' (2013) *Omega* program.

Results

The application of CFA to the 1,000 simulated WISC-V, WJ IV Cognitive, DAS-II, and KABC-II correlation matrices is summarized in Table 1. Results of Table 1 include the number of successful replications along with averaged CFA fit indices (i.e., chi-square, RMSEA, CFI/TLI, SMR, AIC, and BIC). Tables 2 through 5 report the average sources of variances and metrics of interpretability for the models determined to have the best global and local fit. This includes standardized estimates, explained common and total variance, communality/uniqueness, and metrics of interpretability including omega coefficients, PUC, and *H*. As noted in Table 1, none of the structures posited by the publishers emerged as having superior fit to that arrived upon in the independent literature. For clarity, best fitting model results are also summarized in Table 6.

WISC-V

CFA simulation results of the WISC-V primary and secondary 16 subtests are summarized in Table 1 while the best fitting model is presented in Table 2. These results compared the WISC-V technical manual's five factor higher-order structure with several competing models investigated in the independent literature. This included a modified five factor higher-order model (Reynolds & Keith, 2017), a four factor bifactor model (Canivez & Watkins, 2016), and a three factor bifactor model (Dombrowski, Beaujean, Canivez & Watkins, 2015). As noted, the simulation results for the WISC-V revealed that the four bifactor structure [one general factor and four group factors (Gc, PRI, Gsm, & Gs)] reported by Canivez and Watkins (2016) was deemed superior to that of Reynolds and Keith (2017) in terms of global fit statistics and local fit. This resulting structure is presented in Table 2. Nevertheless, this model was only able to converge in 987 of the 1,000 runs in comparison to all of the other models which had a perfect convergence rate. Dombrowski et al.'s (2015) EBFA model produced superior fit statistics to that produced by the test publisher but inferior to that of Canivez and Watkins and Reynolds and Keith.

WJ IV Cognitive

Simulation of the WJ IV Cognitive 18 subtest standard and extended battery is summarized in Table 1 while the best fitting simulated model is presented in Table 3. These results compared the technical manual's seven factor higher-order structure at ages 9 to 13 with Dombrowski et al's (2017, 2018) structure, which was originally guided by both exploratory and confirmatory factor analyses. Simulation results for the WJ IV Cognitive revealed that a four bifactor structure [one general and four group factors (Gc, PR, Gsm, & Gs)] reported by Dombrowski and colleagues produced the best overall fit. Please see Table 3 for standardized coefficients, variance partitioning, and metrics of interpretability. Of concern, the simulated model suggested by the test publisher produced several negative variance estimates and structural path coefficients indicative of model misspecification. All of the models converged in all the runs.

KABC-II

Simulation results of the KABC-II sixteen core and supplemental subtests is furnished in Table 1 while the best fitting simulated model is presented in Table 4. The results compared the technical manual's five factor higher order (Gc, Gf, Gv, Gsm, & Glr) structure with that of McGill and Dombrowski (2018), and that of Reynolds et al. (2007). Simulation results for the KABC-II revealed that a five factor higher-order factor produced by Reynolds et al. (2007) was the preferred model, readily permitted all but two subtests to have primarily loadings on their theoretically proposed factors, and produced the best overall global and local fit. As previously mentioned, this model deviates slightly form the model suggested by the test publisher with additional parameters in the form of a theoretically inconsistent Gestalt Closure loading, subtest cross-loadings, and correlated residual terms between the Glr subtests and related Delayed Recall tasks. All models had a perfect convergence rate.

DAS-II

Simulation results of the DAS-II 20 core and diagnostic subtests is furnished in Table 1 and the best fitting model is presented in Table 5. The results compared the technical manual's simulated six factor higher-order (Gc, Gf, Gv, Gsm, Gs, & Glr) structure, the SL approximate BF structure produced by Dombrowski, McGill, Canivez, & Peterson (2018), and the six factor higher-order model produced by Keith and colleagues (2010). Simulation results for the DAS-II revealed that the six factor higher-order model produced by Keith et al. slightly edged out the model produced by Dombrowski and colleagues in terms of overall fit. Whereas the alternative BF model produced a superior CFI value, the higher-order model yielded better fit, in comparison, on the TFI, RMSEA, and BIC indices. Additionally, whereas the higher-order model converged in all of the runs, the BF model failed to converge in 18 of the runs.

Summary of Loadings, Variance Partitioning, and Metrics of Interpretability

Table 6 summarizes the results for the CFA analyses for all four tests. It is noted that the simulated independent research models were superior to all of the test publisher's proposed theoretical models. With two of the simulated models (e.g., Canivez and Watkins [2016]) with the WISC-V; and Dombrowski, McGill, & Canivez [2017, 2018] with the WJ IV Cognitive) the independent research arrived at very different structural conclusion from that of the test publisher. With the other two models the CHC theoretical structure emerged as viable following model adjusting in the DAS-II and KABC-II although both resulting structures deviated slightly from that originally proposed by the publisher.

Regardless of whether the final, validated structure involved a bifactor or higher-order model, across all of the best fitting models, the general factor accounted for over 29% of the total variance (29.6% to 38.0%) and anywhere from 60.8% to 73.0% of the common variance. The general factor accounted for 4.0% to 58.0% (Median range: 27.0% to 41.5%) of the subtest variance. Across all models, the group factors accounted for a small proportion of the total (0.2% to 6.2%) and common (0.5% to 12.8%) variance. The general and group factors across the four simulated analyses combined to measure 48.0% to 50.50% of the variance, indicating that between 49.5% and 52.0% of the variance was unexplained (i.e., unique variance).

Omega hierarchical values ranged from .793 (WJ IV Cognitive) to .877 (DAS-II). Omega hierarchical subscale was generally insufficiently low for confident clinical interpretation except for the processing speed (Gs) group factors on the WISC-V and WJ IV Cognitive and the Glr factor on the DAS-II. Hancock and Mueller's index of replicability (H) also supported interpretation at the level of the general factor with all values greater than .70. Consideration of PUC along with ECV suggested that the WISC-V and DAS-II are predominantly unidimensional (i.e., PUC and ECV >.70) while the WJ IV Cognitive and KABC-II approached unidimensionality.

Discussion

The results of this study have several implications for understanding the theory and interpretation of widely used commercial ability measures for children and adolescents. As research, in particular structural validity research, on these measures has been used to refine and develop CHC theory (i.e., Schneider & McGrew, 2018), it may also have implications for our understanding of that particular model/theory of human cognitive abilities. First, in all cases the simulated structure of the independent research-derived models appear to be superior to the simulated structures proposed by test publishers. In some cases, such as the KABC-II and DAS-II, simulation of the structure identified in independent research (i.e., Keith et al., 2010; Reynolds et al., 2007) was essentially consistent with, yet offered superior modification statistics relative to, the test publisher's theoretically proposed five factor higher-order CHC structure requiring only nominal amendments to publisher theory.

In other cases, such as the WISC-V and the WJ IV Cognitive, simulation of the respective independent research of Canivez and Watkins (2016) and Dombrowski, McGill, & Canivez (2017, 2018) resulted in a considerably different model from that proposed by the publishers for those instruments. Of concern, several group factors (as suggested by CHC theory) were not located and the positioning of the general factor was different in many of the researcher-derived models. The results on the WJ IV Cognitive are noteworthy given that the instrument is likely to serve as the preeminent reference measure for CHC experimentation and refinement over the course of the next decade.

Whereas publisher theory (and some independent research) suggests that the influence of general intelligence on the manifest variables (MVs) is mediated through the first-order group factors, an alternative bifactor conceptualization where the general factor has direct influence on the MVs provided the best fit to these data in some cases. Although Beaujean (2015) has argued that Carroll favored a bifactor model when developing his three-stratum model (a forerunner to CHC), other scholars (e.g., Keith & Reynolds, 2012) have argued that a bifactor model is not theoretically compatible with modern conceptualizations of intelligence and it is expected that this debate will continue among intelligence scholars for the foreseeable future.

Replication of Posited Factor Structures

Replication is often referred to as the “gold standard” of science and much attention has been focused lately on the reproducibility of research in the psychological sciences (Schmidt & Oh, 2016). Although *direct* replication—where all aspects of a research design are reconstituted—is often difficult, if not impossible, Lilienfeld (2018) argued that construct validation of psychological tests is potentially an area where direct replication of published findings should be expected given researchers often have access to the same datasets and relevant model parameters are routinely disclosed in published works.

In the present study, all but two of the rival models that were explicated had a perfect replication rate across the 1,000 simulation runs and the two outliers (Dombrowski et al., 2019 [DAS-II]; Dombrowski, McGill, & Canivez, 2017, 2018 [WJ IV Cognitive]) were able to be replicated in over 98% of the runs. Although methodologists have expressed concern about the potential replicability of models derived from specification searches (e.g., MacCallum et al., 1992) and the presence of potentially questionable research practices in the CFA validation literature (Beaujean, 2016), these results indicate that the models that are posited for commercial ability measures are able to be replicated and that it may be possible even to replicate complexly determined models when large sample sizes are employed. Nevertheless, Hutchinson (1998) demonstrated that modifications tend to be unstable unless sample sizes are quite large (e.g., > 1,200) and that caution should be exercised in interpreting modified models with smaller samples then the ones employed in this investigation. It is worth noting that the normative samples for most commercial ability measures contain approximately 1,000-2,000 participants.

However, these results also suggest that reliance on a single factor analytic study to ascertain what an intelligence test measures and the structuring of variables for a given measure may be problematic. For example, there are now several posited models for the WISC-V in the professional literature, all of which were found to be replicable in the present study. Yet, when simulated against one another the results favor a more parsimonious model consistent with previous Wechsler Theory (Canivez & Watkins, 2016), even though a major goal of the revision was to align the test better with the CHC model. The discrepancies between publisher theory and the best fitting models identified by independent researchers for all of the tests that were evaluated in present study suggest that structural validation for a given measure is likely to evolve as subsequent independent research emerges in the years after an instrument is published.

Is Model Complexity Desirable: Theoretical Versus Practical Applications?

As previously mentioned, a number of post hoc adjustments were needed in order to discover the best fitting models for the DAS-II and KABC-II. These models contain several complex parameters that depart from desired simple structure. However, from an applied measurement perspective questions remain about how to adapt complexly determined theoretical models to a measurement instrument that will be scored and interpreted by clinical assessment professionals. Although it may be appropriate to engage in a series of model adjustments to help understand theory (i.e., CHC), this approach may be considered less appropriate when an interpretable scale is of importance. When creating instruments for the purpose of theory building, greater leniency toward the practice of model adjustment may well be afforded to the researcher than when attempting to build scales for interpretation. In a traditional reflective model of scale development researchers should guard against secondary loaded indicators, and latent constructs that require correlation of disturbance terms as it is unknown how these indicators and constructs should be scaled to create an interpretable score for a target group specific factor (Reise, 2012). Should the correlated disturbance terms be ignored? Should the indicator be assigned to the factor with the highest loading? Should the indicator be assigned to both factors? These are yet unanswered questions but it could be argued that traditional scale development would favor either eliminating the cross loaded indicator or reconsidering the final, adopted structure if there is the finding of latent group constructs that need significant modifications in order for a permissible solution to converge (Brown, 2015).

For instance, consider the Reynolds and Keith (2017) model that claimed the WISC-V can be interpreted as recommended by the publisher. How would a group factor score be appropriately derived from an instrument that has an indicator (i.e., Arithmetic) load on a latent group factor and a higher-order general factor that also contributes to that same group factor? Although theory building may permit this type of model adjusting we contend that an instrument created primarily for the purpose of clinical assessment should not be so complexly determined. Whereas this complexity may better reflect the true nature of the relations between cognitive abilities (Schneider & McGrew, 2018), it is important to consider that commercial ability measures continue to be constructed in a way that assumes perfect cluster structure.

Higher-Order Versus Bifactor

There is a second interesting line of thinking that emerges from the agglomeration of the results of this study. Although independent research may point to a possible dichotomy when discussing whether a bifactor or higher-order model is the “best” model for conceptualizing the structure of our IQ tests, it appears that this conclusion is more nuanced. Within this simulation study, the best model fit depended upon the specific instrument investigated. With the WISC-V and WJ IV Cognitive the bifactor model appears superior in terms of both global and local fit. With the KABC-II the higher-order model appears to fit the data better than the bifactor model. With the DAS-II simple structure is for the most part attained and the data fit either model fairly well. Thus, the results of this study suggest, from at least a theoretical perspective (and perhaps even from an applied measurement perspective), the preference for a higher-order or bifactor conceptualization of IQ test structure may depend upon context (i.e., the instrument under investigation and the sample being evaluated).

Given concerns in the literature that have been raised about whether better fitting bifactor models are an artifact of the model itself (Murray & Johnson, 2013; Reise, Kim, Mansolf, &

Widaman, 2016) there is not consistent agreement regarding whether the bifactor model represents a viable *theory* for approximating the structure of intelligence. On the one hand, Reynolds and Keith (2013) argued in favor of the higher-order model suggesting that "…higher-order models are theoretically more defensible, more consistent with relevant intelligence theory (e.g., Jensen, 1998), than are less constrained hierarchical [bifactor] models" (p. 66).

On the other hand, several researchers (e.g., Beaujean, 2015; Brunner et al., 2012; Gignac, 2006, 2008; Gignac & Watkins, 2013; Gustafsson & Balke, 1993) have posited a bifactor theoretical conceptualization of intelligence contending that general intelligence is the most substantive factor in a battery of cognitive tests (and subtests) so *g* should be modeled *directly.* In support, these researchers posit that Spearman's (1927) and Carroll's (1993) conceptualization of intelligence is likely better reflected by the bifactor model since the general factor is directly involved in all observed cognitive abilities, not indirectly involved, or mediated by other factors. Thus, this position would argue that it is the higher-order model that requires greater theoretical justification as full mediation of the general factor by the group factors may not reflect the reality of how intelligence operates at least within our IQ tests. Still others have suggested that it might be best to regard the bifactor model as useful primarily for variance partitioning and clarifying how psychological tests should be interpreted in clinical practice (Bonifay, Lane, & Reise, 2017). The debate over whether the bifactor or higher-order model is "superior," or whether the bifactor model is best used as a tool for variance partitioning for purpose of interpretation remains unresolved and in need of further discussion.

Implications for the Clinical Assessment of Intelligence

The next issue to be addressed moves us more directly into the realm of clinical interpretation. Should the group factors be interpreted the way suggested in the respective test publisher's technical manuals? As previously mentioned, two of the simulated instruments located the theoretically proposed CHC group factors (yet not all subtests aligned according to publisher theory) but two did not. An empirically-guided approach would be to cease interpretation of the WISC-V and WJ IV Cognitive the way suggested in their respective technical manuals (or at least eschew the interpretation of specific scores). We cannot have confidence in a publisher's proposed factor structures when subtests do not load on their theoretically proposed group factors. On the other hand, how are we to interpret an instrument when most subtests aligned with their theoretically proposed factors and all theoretically proposed factors emerged as in the case of the KABC-II and DAS-II. Should we just move forward with interpretation basing our decision to do so solely on basis of presumed theoretical alignment (i.e., CHC)?

The answer to this question is that theoretical alignment of subtests with group factors is a necessary but insufficient condition for interpretation. Although alignment of group factors with theory is important, the consideration of variance partitioning and metrics of interpretability is equally important and should be an essential arbiter of the decision to interpret an intelligence test. Regardless of whether one adopts a higher-order or bifactor conceptualization of IQ test structure the results of this simulation suggest that our IQ tests may be confidently interpreted at the general factor level but caution should be exercised when moving to an interpretation at the group factor level. With the WISC-V and WJ IV Cognitive only the Gs-derived index (see Table 6) may be confidently interpreted. With the DAS-II only the Glr index may be interpreted beyond general intelligence. In the KABC-II none of the indices contained sufficient variance for confident clinical interpretation.

It is unknown why the structure posited in the various IQ assessment's technical manuals have worse fit than that of independent research? It is understood that the creation of an IQ test is a costly enterprise. As a result, the potential threat of confirmation should be considered, especially when a preferred theoretical structure is disclosed well in advance of formal model testing (e.g., Weiss, Keith, Zhu, & Chen, 2013). To guard against the effects of sunk costs, test publishers interested in scale development should consider creating a pilot study where the battery is administered to a small group of participants and EFA and CFA is used to establish initial validity evidence. Then MC modeling can be used as a tool to test the replicability of various models. In this way, problematic models may be identified and addressed before commencing with large scale standardization efforts.

Study Limitations

We stipulate that MC simulation is not a panacea for the ills of poor modeling in IQ assessment. The simulation itself should be predicated upon a plausible theoretical conceptualization. If the theory itself is deficient then the resulting MC model based upon that theory could result in the adoption and reification of an erroneous model. Additionally, although the present study employed CHC as a theoretical framework for interpreting factor analytic results, it is important to acknowledge that, despite the attention CHC has received in the psychometric literature over the course of the last 20 years, other plausible models for intelligence structures have been proposed during that same time span. For example, a more parsimonious model adapted from Vernon's (1950) hierarchical model positing verbal, perceptual, and image rotation group factors (VPR) has been found to fit several large cognitive datasets better than CHC (Major, Johnson, & Deary, 2012). Although McGrew (2009) has acknowledged these findings and has called for VPR to be considered in future CHC research, VPR has, thus far, received scant attention in the applied validity literature. Finally, the simulations in the study are based on the normative data. It is not clear whether these results will generalize to clinical populations, and additional research on this topic is necessary.

Conclusion

It is believed that this study will serve as a useful overview, and subsequent replication, of the structure of prominent commercial ability measures. It suggests the inconsistent alignment with CHC theory in two instruments but not in two others. It coheres around the conclusion that our IQ tests, whether theoretically consistent or not, are predominantly measures of general ability and may only nominally measure many of the group factors posited by publisher theory and independent researchers (e.g., Benson, Beaujean, McGill, & Dombrowski, 2018; Carretta & Ree, 2001; Carroll, 1993, 1995; 2003; Cucina & Howardson, 2017). More importantly, it raises concern about omnibus theories of intelligence (i.e., CHC) that have been developed largely on the basis of research summaries. As illustrated in the present study, the CHC factor analytic literature is complex and inconsistent. As conceptualized in modern commercial ability measures, questions remain about the nature and influence of the general factor, the "true" structure of intelligence, how many group factors exist, and whether their measurement can be replicated consistently across tests that are derived from the same theory, and purport to measure the same constructs. To be clear we are not suggesting that CHC theory, as presently constructed, lacks verisimilitude. We are simply encouraging researchers and practitioners to consider these limitations when consulting summaries of the CHC factor-analytic literature and using those resources as a focal point for guiding clinical assessment.

References

Bandalos, D. L., & Leite, W. (2013). Use of Monte Carlo studies in structural equation modeling research. In G. R. Hancock & R. O. Mueller (Eds), *Structural equation modeling: A second course* (2nd ed) (pp. 625-666). Charlotte, NC: Information Age.

Beaujean, A. A. (2015). John Carroll's views on intelligence: Bi-factor versus higher-order models. *Journal of Intelligence, 3 (4),* 121-136. doi: 10.3390/jintelligence3040121

Beaujean, A. A. (2016). Reproducing the Wechsler Intelligence Scale for Children-Fifth Edition: Factor model results. *Journal of Psychoeducational Assessment, 34,* 404-408. Doi: 10.1177/0734282916642679

Benson, N. F., Beaujean, A. A., McGill, R. J., & Dombrowski, S. C. (2018). Revisiting Carroll's survey of factor-analytic studies: Implications for the clinical assessment of intelligence. *Psychological Assessment, 30,* 1028-1038. doi: 10.1037/pas0000556

Bonifay, W., Lane, S. P., & Reise, S. P. (2017). Three concerns with applying a bifactor model as a structure of psychopathology. *Clinical Psychological Science, 5,* 184-186. doi: 10.1177/2167702616657069

Box, G. E. P., & Draper, N. R. (1987). *Empirical model-building and response surfaces*. New York: Wiley.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York: Guilford Press.

Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *Journal of Personality, 80*, 796-846. doi: 10.1111/j.1467-6494.2011.00749.x

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33*, 261-304. doi: 10.1177/0049124104268644

Byrne, B. M. (2005). Factor analytic models: Viewing the structure of an assessment instrument from three perspectives. *Journal of Personality Assessment, 85*, 17-32. doi: 10.1207/s15327752jpa8501_02

Canivez, G. L. (2008). Orthogonal higher-order factor structure of the Stanford-Binet

Intelligence Scales–Fifth Edition for children and adolescents. *School Psychology Quarterly, 23*, 533–541. doi: 10.1037/a0012884

Canivez, G. L., & Watkins, M. W. (2016). Review of the Wechsler Intelligence Scale for Children-Fifth Edition: Critique, commentary, and independent analyses. In A. S. Kaufman, S. E. Raiford, & D. L. Coalson, *Intelligent testing with the WISC-V* (pp. 683–702). Hoboken, NJ: Wiley.

Canivez, G. L., Dombrowski, S. C., & Watkins, M. W. (2018). Factor structure of the WISC-V for four standardization age groups: Exploratory and hierarchical factor analyses with the 16 primary and secondary subtests. *Psychology in the Schools*, *55*, 741-769. doi: 10.1002/pits.22138

Canivez, G. L., McGill, R. J., Dombrowski, S. C., Watkins, M. W., Pritchard, A. E., & Jacobson, L. A. (2018). Construct validity of the WISC-V in clinical cases: Exploratory and confirmatory factor analyses of the 10 primary subtests. *Assessment*. Advance online publication. doi: 10.1177/1073191118811609

Canivez, G. L., Watkins, M. W., & Dombrowski, S. C. (2016). Factor structure of the Wechsler Intelligence Scale for Children-Fifth Edition: Exploratory factor analyses with the 16 primary and secondary subtests. *Psychological Assessment, 28*, 975-986. doi: 10.1037/pas0000238

Carretta, T. R., & Ree, J. J. (2001). Pitfalls of ability research: *International Journal of Selection and Assessment, 9,* 325-335. doi: 10.1111/1468-2389.00184

Carroll, J. B. (1993). *Human cognitive abilities*. New York: Cambridge University Press.

Carroll, J. B. (1995). On methodology in the study of cognitive abilities. *Multivariate Behavioral Research, 30,* 429-452. doi: 10.1207/s15327906mbr3003_6

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

Casey, T. M., & Harden, J. J. (2014). *Monte Carlo simulations and resampling methods for social science*. Thousand Oaks, CA: Sage.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*, 464-504. doi: 10.1080/10705510701301834

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

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

Cucina, J. M., & Howardson, G. N. (2017). Woodcock-Johnson-III, Kaufman Adolescent and Adult Intelligence Test (KAIT), Kaufman Assessment Battery for Children (KABC), and Differential Ability Scales (DAS) Support Carroll but Not Cattell-Horn. Psychological *Assessment, 29,* 1001-1015. doi:10.1037/pas0000389.

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

Dombrowski, S. C., Canivez, G. L., & Watkins, M. W. (2018). Factor structure of the 10 WISC-V primary subtests across four standardization age groups. *Contemporary School Psychology, 22,* 90-104. doi: 10.1007/s40688-017-0125-2

Dombrowski, S. C., Canivez, G. L., Watkins, M. W., & Beaujean, A. (2015). Exploratory bifactor analysis of the Wechsler Intelligence Scale for Children–Fifth Edition with the 16 primary and secondary subtests. *Intelligence*, *53*, 194-201. http://dx.doi.org/10.1016/j.intell.2015.10.009

Dombrowski, S. C., McGill, R. J., & Canivez, G. L. (2017). Exploratory and hierarchical factor analysis of the WJ-IV Cognitive at school age. *Psychological Assessment, 29*, 394-407. doi: 10.1037/pas0000350

Dombrowski, S. C., McGill, R. J., & Canivez, G. L. (2018). An alternative conceptualization of the theoretical structure of the Woodcock-Johnson IV Tests of Cognitive Abilities at school age: A confirmatory factor analytic investigation. *Archives of Scientific Psychology, 6*, 1-13. doi: 10.1037/arc0000039

Dombrowski, S. C., McGill, R. J., Canivez, G. L., & Peterson, C. H. (2019). Investigating the Theoretical Structure of the Differential Ability Scales-Second Edition Through Hierarchical Exploratory Factor Analysis. *Journal of Psychoeducational Assessment, 37*, 91–104. doi: 10.1177/0734282918760724

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*, 494-507. doi: 10.1177/0734282909333179

Elliott, C. D. (2007). *Differential Ability Scales* (2nd ed.). San Antonio, TX: Harcourt Assessment.

Gignac, G. E. (2006). The WAIS–III as a nested factors model: A useful alternative to the more conventional oblique and higher-order models. *Journal of Individual Differences, 27,* 73-86. doi:10.1027/1614-0001.27.2.73

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

Gignac, G. E., & Watkins, M. W. (2013). Bifactor modeling and the estimation of model-based reliability in the WAIS–IV. *Multivariate Behavioral Research, 48,* 639-662. doi:10.1080/00273171.2013.804398

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum

Gustafsson, J. E., & Åberg-Bengtsson, L. (2010). Unidimensionality and interpretability of psychological instruments. In S. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 97-121). Washington, DC: American Psychological Association.

Gustafsson, J.-E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research, 28,* 407-434. doi:10.1207/s15327906mbr2804_2

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. Du Toit & D. Sorbom (Eds.), *Structural equation modeling: Present and future* (pp. 195-216). Lincolnwood, IL: Scientific Software International.

Harwell, M. R., Stone, C. A., Hsu, T-C., & Kirisci, L. (1996). Monte Carlo studies in Item Response Theory. *Applied Psychological Measurement, 20,* 101-125.

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

Hu, L.T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 5,* 1-55. doi: 10.1080/10705519909540118

Hutchinson, S. R. (1998). The stability of post hoc model modifications in confirmatory factor analysis models. *Journal of Experimental Education*, *66,* 361-380. doi: 10.1080/00220979809601406

Jennrich, R. I., & Bentler, P. M. (2011). Exploratory bi-factor analysis. *Psychometrika*, *76,* 537–549. doi:10.1007/s11336–011–9218–4.

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

Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Assessment Battery for Children* (2nd ed.). Circle Pines, MN: American Guidance Service.

Keith, T. Z., & Kranzler, J. H. (1999). The absence of structural fidelity precludes construct validity: Rejoinder to Naglieri on what the Cognitive Assessment System does and does not measure. *School Psychology Review, 28*, 303-321.

Keith, T. Z., Low, J. A., Reynolds, M. R., Patel, P. G., & Ridley, K. P. (2010). Higher-order factor structure of the Differential Ability Scales-II: Consistency across ages 4 to 17. *Psychology in the Schools, 47*, 676–697. doi: 10.1002/pits.20498

Keith, T. Z., & Reynolds, M. R. (2012). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (4th ed., pp. 758-799). New York: Guilford Press.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2,* 196–217. doi: 10.1207/s15327957pspr0203_4

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: Guilford Press.

Lai, K., & Green, S. B. (2016). The problem with having two watches: Assessment of fit when RMSEA and CFI disagree. *Multivariate Behavioral Research, 51,* 220-239. doi: 10.1080/00273171.2015.1134306

Lilienfeld, S. O. (2018). Direct replication and clinical psychological science. *Behavioral and Brain Sciences*. Advance online publication. doi: 10.1017/S0140525X18000754

MacCallum, R. C., Roznowski, M., & Neocowitz, L. B. (1992). Model modifications in covariance structure analysis: the problem of capitalization on chance. *Psychological Bulletin, 111,* 490-504. doi: 10.1037/0033-2909.111.3.490

Major, J. T., Johnson, W., & Deary, I. J. (2012). Comparing models of intelligence in Project TALENT: The VPR model fits better than the CHC and extended Gf-Gc models. *Intelligence, 40,* 543-559. doi: 10.1016/j.intell.2012.07.006

Mansolf, M., & Reise, S. P. (2016). Exploratory bifactor analysis: The Schmid-Leiman orthogonalization and Jennrich-Bentler analytic rotations. *Multivariate Behavioral Research, 51,* 698-717. doi: 10.1080/00273171.2016.1215898

McDonald, R. P. (2010). Structural models and the art of approximation. *Perspectives on Psychological Science, 5,* 675-686. doi: 10.1177/1745691610388766

McGill, R. J., & Dombrowski, S. C. (2018). Factor structure of the CHC model for the KABC-II: Exploratory factor analyses with the 16 core and supplemental subtests. *Contemporary School Psychology, 22,* 279-293. doi: 10.1007/s40688-017-0152-z

McGill, R. J., Dombrowski, S. C., & Canivez, G. L. (2018). Cognitive profile analysis in school psychology: History, issues, and continued concerns. *Journal of School Psychology, 71,* 108-121. http://dx.doi.org/10.1016/j.jsp.2018.10.007

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

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806-834. doi: 10.1037/0022-006X.46.4.806

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66,* 195-244. doi: 10.2466/pr0.1990.66.1.195

Metropolis, N., & Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association, 44,* 335-341. doi: 10.2307/2280232

Murray, A. L., & Johnson, W. (2013). The limitations of model fit comparing the bi-factor versus higher-order models of human cognitive ability structure. *Intelligence, 41,* 407-

422. doi: 10.1016/j.intell.2013.06.004

Muthén, L. K., & Muthén, B. O. (1998 –2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Author.

Powell, D. A., & Schafer, W. D. (2001). The robustness of the likelihood ratio chi-square test for structural equation models: A meta-analysis. *Journal of Educational and Behavioral Statistics, 26,* 105-132.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*, 667-696. doi: 10.1080/00273171.2012.715555

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment, 95*, 129-140. doi: 10.1080/00223891.2012.725437

Reise, S. P., Kim, D. S., Mansolf, M., & Widaman, K. F. (2016). Is the bifactor model a better model or is it just better at modeling implausible responses? Application of iteratively reweighted least squares to the Rosenberg Self-esteem Scale. *Multivariate Behavioral Research, 51,* 818-838. doi: 10.1080/00273171.2016.1243461

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

Reynolds, M. R. & Keith, T. Z. (2017). Multi-group and hierarchical confirmatory factor analysis of the Wechsler Intelligence Scale for Children-Fifth Edition: What does it measure? *Intelligence, 62,* 31-47. doi: 10.1016/j.intell.2017.02.005.

Reynolds, M. R., Keith, T. Z., Fine, J. G., Fisher, M. E., & Low, J. (2007). Confirmatory factor structure of the Kaufman Assessment Battery for Children-Second Edition: Consistency with Cattell-Horn-Carroll theory. *School Psychology Quarterly, 22*, 511-539. doi:10.1037/1045-3830.22.4.511

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*, 137-150. doi: 10.1037/met0000045

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions.

Psychometrika, 22, 53-61. doi: 10.1007/BF02289209

Schmidt, F. L., & Oh, I. (2016). The crisis of confidence in research findings in psychology: Is lack of replication the real problem? Or is it something else? *Archives of Scientific Psychology, 4 (1)*, 32-37. doi: 10.1037/arc0000029

Schneider, W. J., & McGrew, K. S. (2018). The Cattell-Horn-Carroll Theory of Cognitive Abilities. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (4th ed., pp. 73-163). New York: Guilford Press.

Schrank, F. A., McGrew, K. S., & Mather, N. (2014). *Woodcock–Johnson IV Tests of Cognitive Abilities*. Rolling Meadows, IL: Riverside.

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology, 69,* 487-510. doi: 10.1146/annurev-psych-122216-011845

Spearman, C. (1927). *The abilities of man*. New York: Cambridge.

Sotelo-Dynega, M. and Dixon, S. G. (2014), Cognitive assessment practices: A survey of school psychologists. *Psychology in the Schools, 51*, 1031-1045. doi: 10.1002/pits.21802

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

Watkins, M. W. (2013). Omega [Computer software]. Phoenix, AZ: Ed & Psych Associates.

Watkins, M. W. (2017). The reliability of multidimensional neuropsychological measures: From alpha to omega. *The Clinical Neuropsychologist*, *31*, 1113-1126. doi: 10.1080/13854046.2017.1317364

Watkins, M. W., Dombrowski, S. C., & Canivez, G. L. (2018). Reliability and factorial validity of the Canadian Wechsler Intelligence Scale for Children–Fifth Edition. International *Journal of School and Educational Psychology*, *6*, 252–265. doi: 10.1080/21683603.2017.1342580

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children* (4th ed.). San Antonio, TX: Psychological Corporation

Wechsler, D. (2014). *Wechsler Intelligence Scale for Children* (5th ed.). San Antonio, TX: NCS Pearson.

Weiss, L. G., Keith, T. Z., Zhu, J., & Chen, H. (2013). WISC-IV and clinical validation of the four- and five-factor interpretative approaches. *Journal of Psychoeducational Assessment, 31,* 114-131. doi: 10.1177/0734282913478032

Table1

Monte Carlo Simulation of Posited Models for Prominent Commercial Ability Measures (Average Fit Statistics Across 1,000 Replications)

Model	χ^2	*df*	CFI	TLI	SRMR	RMSEA	BIC	AIC	Replications
			WISC-V						
BF (Canivez & Watkins, 2016)	**401.8**	**88**	**0.980**	**0.973**	**0.024**	**0.040**	**84982**	**84617**	**987**
HO (Reynolds & Keith , 2017)	437.7	97	0.978	0.973	0.026	0.040	84948	84635	1000
HO (Publisher Theory)	1095.4	98	0.936	0.922	0.108	0.068	85598	85290	1000
BF (Dombrowski et al., 2015)	697.3	93	0.961	0.95	0.038	0.054	85238	84902	1000
			WJ IV Cognitive						
BF (Dombrowski, McGill & Canivez, 2017, 2018)	**1669.3**	**118**	**0.867**	**0.827**	**0.047**	**0.091**	**70657**	**70277**	**1000**
HO Publisher Theory[a]	2046.4	128	0.835	0.803	0.056	0.098	70961	70634	1000
			KABC-II						
HO (Reynolds et al., 2007)	**530.0**	**95**	**0.971**	**0.964**	**0.029**	**0.047**	**77515**	**77195**	**1000**
BF (McGill & Dombrowski, 2017)	661.4	88	0.962	0.949	0.036	0.057	77667	77307	1000
HO (Publisher Theory)	1437.6	97	0.912	0.891	0.046	0.083	78406	78097	1000
			DAS-II						
HO (Keith et al., 2010)	**522.5**	**161**	**0.949**	**0.940**	**0.035**	**0.053**	**38320**	**37998**	**1000**
BF (Dombrowski, McGill & Canivez, & Peterson, 2018)	509.9	155	0.951	0.939	0.036	0.054	38348	37998	982
HO (Publisher Theory)	616.0	164	0.936	0.926	0.040	0.059	38394	38085	1000

Note. CFI = comparative fit index; TLI = Tucker-Lewis Index; SRMR = standardized root mean square; RMSEA = root mean square error of approximation; BIC = Bayesian Information Criterion; AIC = Akaike's Information Criterion. BF = bifactor; HO = higher-order.

[a]WJ IV Cog publisher theory reported negative residual variance of -2.29 and -0.043 on Ga and a standardized parameter estimate of *g* on Ga of 1.13. **Bold** text illustrates best fitting model.

Table 2

Average Sources of Variance for WISC-V According to a Bifactor Model (Canivez & Watkins, 2016)

	General		Verbal Comprehension (Gc)		Perceptual Reasoning (Gf/Gv)		Working Memory (Gsm)		Processing Speed (Gs)			
Subtest	*b*	S^2	*b*	S^2	*b*	S^2	*b*	S^2	*b*	S^2	h^2	u^2
Similarities (Gc)	.72	.52	.35	.12							.64	.36
Vocabulary (Gc)	.73	.53	.47	.22							.74	.26
Information (Gc)	.72	.52	.38	.15							.67	.33
Comprehension (Gc)	.63	.39	.32	.10							.50	.50
Block Design (Gv)	.64	.41			.38	.14					.56	.45
Visual Puzzles (Gv)	.65	.42			.50	.25					.69	.32
Matrix Reasoning (Gf)	.64	.41			.13	.02					.43	.57
Figure Weights (Gf)	.65	.42			.16	.03					.45	.55
Picture Concepts (Gf)	.53	.28			.06	.00					.29	.71
Arithmetic (Gsm/Gf/Gc)	.74	.54					.49	.24			.56	.44
Digit Span (Gsm)	.66	.44					.13	.02			.68	.32
Picture Span (Gsm)	.55	.30					.30	.09			.39	.61
Letter–Number Seq. (Gsm)	.65	.42					.45	.20			.63	.37
Coding (Gs)	.37	.13							.63	.39	.53	.47
Symbol Search (Gs)	.43	.18							.68	.46	.64	.36
Cancellation (Gs)	.19	.04							.37	.14	.17	.83
Total Variance		.372		.037		.027		.035		.062	.48	.52
Common Variance		.696		.069		.051		.065		.115		
ω_H/ω_{HS}		.844		.302		.110		.181		.518		
H		.915		.567		.354		.407		.626		
PUC		.792										

Note. *b* = standardized loading of subtest on factor; S^2 = variance explained in the subtest; h^2 = communality; u^2 = uniqueness; ω_H = Omega-hierarchical (general factor), ω_{HS} = Omega-hierarchical subscale (group factors), H = construct reliability or replicability index, PUC = percentage of uncontaminated correlations. Posited group factor alignment in parentheses.

Table 3

Average Sources of WJ IV Variance According to a Bifactor Model (Dombrowski et al., 2017, 2018)

	g		Working Memory (Gsm)		Perceptual Reasoning		Processing Speed (Gs)		Verbal Ability (Gc)			
Subtest	*b*	S^2	*b*	S^2	*b*	S^2	*b*	S^2	*b*	S^2	h^2	u^2
Verbal Attention (Gwm)	.59	.35	.44	.20							.55	.45
Memory for Words (Aud Mem)	.52	.27	.49	.24							.51	.49
Object Number Sequence (Gsm)	.63	.39	.36	.13							.52	.48
Nonword Repetition (Ga)	.46	.21	.36	.13							.35	.65
Phonological Processing (Ga)	.62	.38	.24	.06							.44	.56
Numbers Reversed (Gsm)	.57	.33	.16	.03							.36	.64
Visualization (Gv)	.51	.26			.50	.25					.51	.49
Visual-Auditory Learning (Glr)	.43	.19			.38	.14					.34	.67
Picture Recognition (Gv)	.35	.12			.43	.18					.31	.69
Analysis-Synthesis (Gf)	.62	.38			.31	.10					.48	.52
Concept Formation (Gf)	.63	.39			.21	.05					.44	.56
Story Recall (Glr)	.52	.27			.19	.04					.30	.70
Letter-Pattern Matching (Gs)	.50	.25					.60	.36			.62	.38
Number-Pattern Matching (PerSpd)	.49	.24					.60	.36			.61	.39
Pair Cancellation (Gs)	.41	.17					.61	.37			.55	.45
Number Series (Gf)	.66	.43					.13	.02			.45	.55
Oral Vocabulary (Gc)	.67	.45							.61	.37	.83	.17
General Information (Gc)	.50	.25							.61	.37	.62	.38
Total Variance		.296		.043		.042		.062		.042	.487	.513
Common Variance		.608		.088		.086		.128		.085		
ω_H/ω_{HS}		.793		.221		.239		.525		.436		
H		.890		.485		.485		.695		.542		
PUC		.768										

Note. *b* = factor loading, S^2= variance explained, h^2= communality, u^2= uniqueness, ω_H = Omega hierarchical (*g*), ω_{HS} = Omega hierarchical subscale (group factors). H =construct replicability. Posited group factor alignment in parentheses.

Table 4

Average Sources of KABC-II Variance According to a Higher-Order Model (Reynolds et al., 2007)

					Residualized				Residualized				Residualized				Residualized				Residualized			
	g	S^2	Gc	S^2	Gc	S^2	Gv	S^2	Gv	S^2	Gf	S^2	Gf	S^2	Glr	S^2	Glr	S^2	Gsm	S^2	Gsm	S^2	h^2	u^2
Verb. Knowledge (Gc)	.70	.49	.84	.71	.47	.22																	.71	.29
Riddles (Gc)	.72	.53	.87	.76	.49	.24																	.76	.24
Ex. Vocabulary (Gc)	.69	.48	.83	.69	.46	.22																	.69	.31
Gestalt Closure (Gv)	.39	.15	.47	.22	.26	.07																	.22	.78
Triangles (Gv)	.62	.38					.72	.51	.36	.13													.51	.49
Rover (Gv)	.51	.26					.59	.34	.30	.09													.34	.66
Block Counting (Gv)	.52	.28					.61	.37	.30	.09													.37	.63
Pattern Reasoning (Gf)	.34	.12					.36	.13	.64	.41	.40	.16	.64	.41									.52	.48
Story Completion (Gf)	.62	.39									.65	.42	.18	.03									.42	.58
Hand Mov. (Gsm)	.34	.12									.36	.13	.47	.22					.29	.08	.47	.22	.34	.66
Atlantis Delayed (Glr)	.44	.19													.55	.30	.33	.11					.30	.70
Rebus Delayed (Glr)	.62	.39													.78	.61	.47	.23					.62	.38
Atlantis (Glr)	.52	.27													.65	.43	.39	.15					.43	.58
Rebus (Glr)	.65	.42													.82	.66	.49	.24					.66	.34
Number Recall (Gsm)	.44	.19																	.69	.48	.54	.29	.48	.52
Word Order (Gsm)	.51	.26																	.81	.66	.63	.40	.66	.34
Total Variance		.308				.046				.019				.039				.046				.043	.502	.498
ECV		.614				.092				.039				.078				.091				.085		
ω_H/ω_{HS}		.822				.264				.170				.328				.288				.265		
H		.888				.485				.257				.503				.474				.517		
PUC		.842																						
Second Order Loadings																								
Gc	.830																							
Gv	.864																							
Gf	.960																							
Glr	.797																							
Gwm	.631																							

Note. Residualized using the following formula: $\sqrt{R^2 - (g\ loading)^2}$. For Hand Movements and Pattern Reasoning the indicator with the higher loading was used to determine variance calculations and omega estimates. Correlated residuals: Atlantis with Atlantis Delayed and Rebus with Rebus Delayed. S^2= variance explained, h^2= communality, u^2= uniqueness, Posited group factor alignment in parentheses.

Table 5

Average Sources of DAS-II Variance According to a Higher-Order Model (Keith et al., 2010)

					Residualized				Residualized				Residualized				Residualized				Residualized				Residualized		
	g	*S*²	Gc	*S*²	Gc	*S*²	Gv	*S*²	Gv	*S*²	Gf	*S*²	Gf	*S*²	Glr	*S*²	Glr	*S*²	Gsm	*S*²	Gsm	*S*²	Gs	*S*²	Gs	*S*²	*h*²
NV	.63	.40	.74	.55	.39	.15																					.55
WD	.62	.38	.73	.53	.38	.15																					.53
VS	.67	.45	.79	.62	.41	.17																					.62
VC	.39	.15	.46	.21	.57	.32					.27	.07	.33	.11													.47
RD	.64	.41					.72	.52	.33	.11																	.52
CP	.58	.34					.66	.43	.30	.09																	.43
LL	.60	.36					.67	.45	.31	.10																	.45
PA	.70	.49					.79	.62	.36	.13																	.62
RP	.49	.24					.55	.30	.25	.06																	.30
MA	.63	.39									.69	.47	.33	.11													.48
SQ	.72	.52									.79	.63	.30	.09													.63
EN	.66	.44									.73	.53	.30	.09													.53
PS	.50	.25									.55	.30	.23	.05													.30
OI	.51	.26													.89	.79	.73	.53									.79
OD	.43	.19													.75	.56	.62	.38									.57
DB	.71	.50																	.74	.55	.21	.04					.55
DF	.64	.41																	.67	.44	.19	.03					.44
SO	.75	.56																	.78	.61	.21	.05					.61
IP	.42	.18																					.56	.31	.36	.13	.31
RN	.48	.23																					.64	.40	.42	.17	.41
Total Var.		.357				.040				.025				.017				.045				.006				.015	.505
ECV		.708				.078				.049				.033				.090				.012				.030	
ω_H/ω_{HS}		.877				.299				.170				.139				.545				.060				.225	
H		.924				.508				.352				.273				.638				.115				.266	
PUC		.858																									

Second Order Loadings

Gc	.851
Gv	.887
Gf	.912
Glr	.573
Gsm	.961

Gs	.759

Note. CP = Copying; DB = Digits Backward, DF = Digits Forward, EN = Early Number Concepts, IP = Speed of Information Processing, LL = Matching Letter Like Forms, MA = Matrices, NV = Naming Vocabulary, OD = Recall of Objects-Delayed, OI = Recall of Objects-Immediate, PA = Pattern Construction-Alternate, PS = Picture Similarities, RD = Recall of Designs, RN = Rapid Naming, SO = Recall of Sequential Order, SQ = Sequential & Quantitative Reasoning, VC = Verbal Comprehension, VS = Verbal Similarities, WD = Word Definitions, RP = Recognition of Pictures. S^2 = variance explained; h^2 = communality; ω_{H} = omega hierarchical; ω_{HS} = omerga hierarchical subscale. Residualized using the following formula: $\sqrt{R^2 - (g\ loading)^2}$. For omega and variance calculations placed VC on Gc. Correlated disturbance: Gf with Gv; Copying with Recall of Designs. Verbal Comprehension crossloads on Gf.

Table 6

Summary of variance partitioning and metrics of interpretability

	WISC-V	WJ IV Cognitive	KABC-II	DAS-II
Best Fitting Model	Bifactor (Canivez & Watkins, 2016).	Bifactor (Dombrowski et al., 2017, 2018)	Higher Order (Reynolds et al., 2007)	Higher Order (Keith et al., 2010)
General *(g)*				
Total S^2 (%)	37.26	29.6	30.8	35.7
Common S^2 (%)	69.6	60.8	61.4	70.8
Subtest S^2 (%)	4.0 to 54.0 Median=41.5	12.0 to 45.0 Median=27.0	.18 to 52.0 Median=27.0	17.0 to 58.0 Median=39.0
g Loadings	.19 to .74	.35 to .67	.34 to .72	.39 to .75
Subtests with poor g loadings	Coding,, Symbol Search & Cancellation	Nonword rep, Vis Aud Lrning, Pic Rep, Num Pattern Match, Pair Cancellation	Gest Closure, Pattern Reasoning, Hand Move, Atlantis Delayed & Number Recall	Verbal Comp., Rec of Pictures, Recall Objects—Delayed, Speed of Info Processing; Rapid Naming
Group				
Total S^2 (%)	2.7 to 6.2	4.2 to 6.2	1.9 to 4.6	.0.6 to 4.0
Common S^2 (%)	5.1 to 11.5	8.5 to 12.8	3.9 to 9.2	1.2 to 9.0
h^2	48.0	48.7	50.2	50.5
u^2	52.0	51.3	49.8	49.5
ω_H	.844	.793	.822	.877
ω_{HS}	.110 to .518	..221 to .525	.170 to .328	.06 to .545
H general/H subtest	.915/.407 to .626	.890/.485 to .695	.888/.257 to .517	.924/.115 to .638
PUC	.792	.768	.842	.858
Interpret Indices Confidently based on metrics of interpretability	Processing Speed (Gs)	Processing Speed (Gs)	None	Glr

Contains all subtests suggested by publisher				
Theory	Gc, Gsm & Gs	Gc	Gsm & Glr	Gc, Gv, Gs Glr & Gsm
Subtest departure from publisher's theory	Arithmetic and all Gf & Gv subtests	9 of 18 subtests	Gestalt Closure on Gc not Gv; Hand Movements and Pattern Reasoning cross load	Verbal Comprehension cross loads Gf

Note. S^2 = variance explained, h^2 = communality, u^2 = uniqueness, ω_H = Omega-hierarchical (general factor), ω_{HS} = Omega-hierarchical subscale (group factors), H = construct reliability or replicability index, PUC = percentage of uncontaminated correlations.