Journal of School Psychology 110 (2025) 101432

Contents lists available at ScienceDirect

Journal of School Psychology

journal homepage: www.elsevier.com/locate/jschpsyc

Conjectures and refutations in cognitive ability structural validity research: Insights from Bayesian structural equation modeling

Stefan C. Dombrowski [a,*], Ryan J. McGill [e], Gary L. Canivez [b], Marley W. Watkins [c], Alison E. Pritchard [d], Lisa A. Jacobson [d]

[a] Rider University, United States of America
[b] Eastern Illinois University, United States of America
[c] Baylor University, United States of America
[d] Kennedy Krieger Institute, Johns Hopkins School of Medicine, United States of America
[e] College of William & Mary, United States of America

ARTICLE INFO

Editor: Andy Garbacz

Keywords:
Bayesian structural equation modeling
WISC–V
Factor analysis
Replication crisis
Bifactor model
Higher order model
Intelligence test

ABSTRACT

The use of Bayesian structural equation modeling (BSEM) provided additional insight into the WISC–V theoretical structure beyond that offered by traditional factor analytic approaches (e.g., exploratory factor analysis and maximum likelihood confirmatory factor analysis) through the specification of all cross loadings and correlated residual terms. The results indicated that a five-factor higher-order model with a correlated residual between the Visual-Spatial and Fluid Reasoning group factors provided a superior fit to the four bifactor model that has been preferred in prior research. There were no other statistically significant correlated residual terms or cross loadings in the measurement model. The results further suggest that the WISC–V ten subtest primary battery readily attains simple structure and its index level scores may be interpreted as suggested in the WISC–V's scoring and interpretive manual. Moreover, BSEM may help to advance IQ theory by providing contemporary intelligence researchers with a novel tool to explore complex interrelationships among cognitive abilities—relationships that traditional structural equation modeling methods may overlook. It can also help attenuate the replication crises in school psychology within the area of cognitive assessment structural validity research through systematic evaluation of complex structural relationships obviating the need for CFA based post hoc specification searches which can be prone to confirmation bias and capitalization on chance.

1. Introduction

Assessment researchers in school psychology may not be familiar with the extension of Bayes' Theorem to structural equation modeling (van de Schoot et al., 2017). This is evidenced by only a single study using this technique appearing in any school psychology journal to date (e.g., Dombrowski et al., 2018). Despite this lacuna, Bayesian structural equation modeling (BSEM) holds promise as a technique for evaluating the latent structure of assessment instruments, especially tests of cognitive ability, within many fields

* Corresponding author at: Rider University, Department of Graduate Education, Leadership, and Counseling, 2083 Lawrenceville Road, Lawrenceville, NJ 08648, United States of America.

E-mail address: sdombrowski@rider.edu (S.C. Dombrowski).

https://doi.org/10.1016/j.jsp.2025.101432
Received 20 December 2023; Received in revised form 26 February 2025; Accepted 26 February 2025
Available online 3 April 2025

including school psychology (Muthén & Asparouhov, 2012). It augurs to overcome some of the limitations of traditional (i.e., frequentist) factor analytic procedures (Brown, 2015) and portends to better reflect not only the measurement model of an instrument but also its underlying theory given recent speculation that the models for some commercial ability measures may be too complex to be evaluated by traditional structural validity methods (e.g., McGrew et al., 2023).

Historically, factor analysis has been regarded as having two major classes: exploratory (or unconstrained factor analysis) and confirmatory (or constrained factor analysis; Gorsuch, 1983). Regardless of the nomenclature involved, no technique is inherently confirmatory or exploratory (Loehlin & Beaujean, 2017). Each class of frequentist factor analysis contains limitations. A potential limitation of traditional confirmatory factor analytic (CFA) estimation is the need to apply overly strict constraints to represent hypotheses about latent structures that are often speculative unless the discrepancy between the researcher's understanding of, and the actual state of, nature equals or approaches zero. This manifests in the general requirement to fix cross-loadings to zero and estimate only a limited number of residual correlations due to available degrees of freedom. When too many parameters are freely estimated using CFA then this can contribute to statistical non-identification and lack of model convergence. CFA can also be prone to the practice of post hoc model modifications that may capitalize on chance (MacCallum et al., 1992; Marsh et al., 2009) and misrepresent the underlying factor structure and theory for an instrument (see also Canivez & Kush, 2013). Over constraining of models can lead to additional problems. For example, when a model is overly constrained, it may overlook important structural relationships that may not have been considered in advance resembling a problem akin to specification bias in standard multiple regression.

On the other hand, exploratory factor analysis (EFA) can partially overcome this limitation by freely estimating all primary and secondary loadings; however, EFA has its own set of limitations. The procedure reduces researcher choice to a decision about the number of factors to extract and retain to arrive at desired simple structure, although there are a variety of methods available to suggest the number of viable factors to extract (Watkins, 2018). Unlike with CFA, EFA does not determine a priori where the indicators will load; instead, loadings are permitted to "speak for themselves" with the factor analytic algorithm determining the location of primary loadings and cross-loadings, if present (Carroll, 1985; Gorsuch, 1983). From a scientific perspective, this may be regarded as a positive feature of EFA especially when less is known about an instrument or its underlying theoretical structure. EFA can work in a complementary fashion with CFA by establishing a baseline model for a new or newly revised instrument against which other models may be tested.

BSEM represents a hybrid of EFA and CFA incorporating aspects of both procedures, and potentially augmenting both with the additive capability of specifying all correlated residuals at the indicator and latent structural level. Since BSEM specifies cross-loadings and correlated residuals using priors that approach, but are not fixed at, zero it may permit an otherwise non-identified model in frequentist CFA to achieve statistical identification. For instance, simple structure in CFA hypothesizes that a variable (i.e., subtest) loads one and only one primary factor (i.e., no crossing loadings). In an IQ test this may not reflect well the true structural reality of the instrument. A more realistic hypothesis might be that selected variables have a major loading on hypothesized factors as well as small cross-loadings due to a minor influence on the variable from some of the other factors. Analogously, some residuals may be correlated because of omission of minor factors or because they share a source of variance unrelated to the general factor. Within the context of CFA, it is exceptionally difficult to envision which residuals should be correlated. Freeing all of these residual variances would lead to a nonidentified model in conventional ML CFA. BSEM offers a possible solution to this problem as many variables in psychology and education are known to be correlated. It is this high intercorrelation among variables within tests of cognitive ability that supports need for the extraction of a higher-order dimension (Thompson, 2004). The higher-order dimension is thought to be something called *g* or general intelligence. In a bifactor model the general factor directly influences all of the subtests (indicators). The general factor indirectly influences subtest variance via mediation through the latent first-orders factors in the higher-order model (please see Canivez [2016], Carroll [1995] and the addendum in the online supplement (https://osf.io/by28n/?view_only=fb977f1304fa48d1adb0798450be6f47).

for a more thorough explication of the comparison of higher-order and bifactor models). Specification of all cross loadings and correlated residual terms could increase clarity of an instrument's underlying structure and connection to theory (Asparouhov & Muthén, 2015; Muthén & Asparouhov, 2012) by arriving at final model in a more efficient manner (i.e., without need for first running EFA followed by CFA with multiple post hoc model specifications). A summary of frequentist EFA and CFA compared to Bayesian structural equation modeling is shown in Table 1. The process of Bayesian estimation is next described.

2. Bayesian estimation

One of the more important considerations when undertaking a Bayesian analysis is the selection of priors. In BSEM, parameters are viewed as variables instead of constants and use a distribution known as a *prior* (Muthén & Asparouhov, 2012; Zyphur & Oswald, 2015). The selection of the prior is important. It is influenced by "prior knowledge" which may be predicated upon theory, pilot studies, exploratory factor analyses, and extant empirical literature (Gelman et al., 2004; Stone, 2013). With Bayesian estimation, the observed data provide information, which is subsequently used to modify a prior into a posterior distribution that produces a median estimate bracketed by a credibility interval. Bayesian estimation produces three different distributions: the prior, the posterior, and the likelihood (Gelman et al., 1996; Gelman et al., 2004). The likelihood represents the distribution of data predicated upon a parameter value. The posterior distribution contains estimated parameter values that fall between the likelihood and the prior. Priors are categorized as either noninformative (i.e., diffuse) or informative. A noninformative prior typically contains a normal distribution with a large variance. When the prior variance is large, the likelihood contributes more information to the formation of the posterior resulting in an estimate closer to the maximum likelihood estimate. When using BSEM this will generally lead to model rejection (Muthén & Asparouhov, 2012) whereby the posterior predictive value hovers around zero.

Table 1
Summary of ML CFA, EFA and BSEM Characteristics.

Characteristics	ML CFA	Traditional EFA	BSEM
Theory	Frequentist	Frequentist	Bayes
Parameters	Constants with standard errors bracketed by a confidence interval	Constants with standard errors bracketed by a confidence interval	Variables with distributions bracketed by a credibility interval
Cross-Loadings	Exact zeros though can be specified requiring a degree of freedom	Freely estimated	Estimated via informative priors (zero mean and small variance)
Major Loading	Freely estimated	Freely estimated	Diffuse noninformative priors (zero mean and infinite variance)
Correlated Residuals	Specified requiring a degree of freedom	Not available	Informative priors (df*D, df) where df = degree of freedom and D = residual variance
Model Modification	Multiple indices with improvement made in a stepwise fashion	Typically, not used but some are available.	All parameters freed and simultaneously estimated. Use of Deviance Information Criteria (DIC) in Mplus and other indices (e.g., Bayes RMSEA) in other statistical applications such as R and WinBugs.
Parameter Estimates	Typically assumed to be normally distributed (not all cases)	Typically assumed to be normally distributed (not in all cases)	Does not assume a normal distribution
Sample Size	Requires large sample size	Requires large sample size	Does not need large samples. With small sample sizes the prior dominates decreasing variance and increasing bias. With larger samples sizes the influence on the posterior is diminished producing estimates closer to those produced by ML CFA causing PPP to not escape from zero.

Note. ML CFA = maximum likelihood confirmatory factor analysis, EFA = exploratory factor analysis, and BSEM = Bayesian structural equation modeling. Adapted from Dombrowski et al. (2018).

2.1. Markov Chain Monte Carlo (MCMC)

Bayesian estimation utilizes MCMC (Edwards, 2010; Green, 1995; Link & Eaton, 2012) algorithms to draw random samples iteratively from the posterior distribution of the model parameters. This data generation process is similar to conventional Monte Carlo simulation, which also utilizes a random sampling technique. The Gibbs algorithm (Casella & George, 1992) is the most popular approach to MCMC sampling. The MCMC algorithm is evaluated for convergence by monitoring the potential scale reduction (PSR) convergence criterion (Gelman et al., 2004; Gelman & Rubin, 1992). The first half of the MCMC chains (i.e., the burn in phase) is used to calibrate the model. The second half of the MCMC chain is used to estimate the posterior distribution (Muthén & Asparouhov, 2012). The PSR criterion compares within- and between-chain variation of parameter estimates. A resulting PSR less than 1.10 indicates an acceptable convergence level while a PSR greater than 1.10 indicates that the model should be rejected. A PSR equivalent to 1.00 is considered perfect model convergence (Kaplan & Depaoli, 2013). Trace and autocorrelation plots may also be evaluated for each parameter to determine model convergence. If plots display a lack of rapid up-and-down fluctuations and an absence of trends over time then the model is considered to have converged properly (Asparouhov & Muthén, 2010; Kaplan & Depaoli, 2013). If a model does not converge then the number of iterations (I) should be increased first by two (2I) and then by four (4I; Muthén & Asparouhov, 2012) until the model attains convergence. An iteration sensitivity analysis is also recommended to determine stability of parameter estimates. However, going beyond 250,000 iterations without convergence likely suggests a model should be rejected. Upon attainment of model convergence, the next step involves an investigation of model fit with the data and consideration of which model might be preferred.

2.2. Model fit and comparison

2.2.1. Posterior predictive checking

Posterior predictive checking is used to determine a model's fit with data. Although research investigating the factor structure of instruments has used the posterior predictive *P*-value (PPP) as a model comparison tool (Cain & Zhang, 2019), it is most appropriately used for checking whether the observed data are similar to the modeled data (Gelman et al., 1996). PPP values range from 0 to 1, with a value of .50 considered perfect model fit (Gelman et al., 1996; Muthén & Asparouhov, 2012). Values of less than .10, or greater than .90 indicate poor model fit with data. The distribution of PPP values is uniform between 0 and 1 (Gelman et al., 1996). In practice, PPP values between .10 and .90 are considered almost equally likely under the null hypothesis. The PPP signals something is wrong with a model when a PPP estimate is at an extreme tail (e.g., $<$.10 or $>$.90). For example, when a PPP is less than .10 or greater than .90 then it should be concluded that the data are not very consistent with the model and the model should be rejected. The PPP may also be assessed for model fit with the data by visually inspecting posterior distribution scatterplots and distribution plots (Muthén & Muthén, 1998–2017). If the scatterplot shows a similar proportion above and below the 45-degree line and the distribution plot demonstrates a balance on both sides of the median line, then the data is considered to have fit the data well.

2.2.2. Deviance information criteria

A researcher may wish to invoke model comparison tools to determine which model is 'superior' or 'best.' These model comparison tools may include the deviance information criterion (DIC; Vehtari & Gelman, 2017), leave-one-out cross-validation (LOO), Bayesian root mean square error of approximation (Hoofs et al., 2018), and the widely applicable information criterion (WAIC). Although available in **R** (R Core Team, 2023), WAIC and LOO are less frequently utilized by researchers because of programming and computational complexity. BRMSEA is available via hand calculation, but validation studies are needed. In Mplus, this generally leaves one model fit index, the DIC, to determine which model is to be preferred (Muthén & Asparouhov, 2012). The DIC is interpreted in the same way as frequentist ML CFA information criterion fit statistics (i.e., AIC and BIC) where lower values are generally preferred although theoretical alignment must also be considered. Of consequence, with BSEM, the need to rely upon multiple modification indices, as occurs in ML CFA, for model respecification may be obviated by the simultaneous estimation of all cross-loadings and correlated error terms. Because all relationships among indicators and factors are estimated simultaneously, this tends to eliminate much of the need for post hoc specification searches that may capitalize on chance.

2.3. Utility of BSEM for tests of cognitive ability and purpose of the study

BSEM may be an especially appropriate methodology for use with instruments that presume to measure correlated traits such as commercial tests of intellectual functioning that often have overlapping constructs left un-evaluated by traditional EFA and CFA techniques. BSEM has been used twice previously to evaluate the structure of cognitive ability measures (e.g., DAS-II [Dombrowski et al., 2018] and WISC-IV [Golay et al., 2013]). Both studies offered additional insight into the factor structure of the respective cognitive ability instruments not previously discussed in the extant, frequentist literature. For instance, Golay et al. (2013) discovered that a direct hierarchical (bifactor) five-factor Cattell-Horn-Carroll (CHC) structure for the WISC–IV (Wechsler, 2003) was superior to the publisher posited four-factor higher-order structure that cohered with prior Wechsler Theory. Dombrowski et al. (2018) found that a two-factor structure for the DAS–II with two subtests only loading on the general factor was superior to the publisher proposed three-factor structure. However, Golay et al. did not employ correlated residuals in their analyses. Dombrowski et al. investigated correlated residuals but not at the latent group factor level.

Accordingly, the present study sought to expand upon the use of BSEM to the WISC–V by applying all features of BSEM technology not previously used in the prior studies (e.g., simultaneous estimation of small variance cross-loadings and correlated residuals for both subtests and group factors). Since its publication, the Wechsler Intelligence Scale for Children-Fifth Edition (WISC–V; Wechsler, 2014a) has been the subject of considerable debate in the empirical literature. Questions remain regarding its *true* underlying factor structure as numerous rival models have been posited based on re-analysis of the normative and clinical sample data. Attempts at replicating the five-factor higher-order structure presented in the manual have been generally unsuccessful. Instead, frequentist EFA and CFA methodologies have suggested an alternative four factor bifactor structure (e.g., Canivez et al., 2017; Dombrowski et al., 2017). Therefore, this study may prove to be a useful replication attempt that provides further insight to help to resolve the debate in the field regarding the theoretical structure of the WISC–V, which is critical given the frequency of its use in school psychology (Benson et al., 2019) and clinical practice. This study may also provide further insight into how the WISC–V should be scored and interpreted given that the scoring structure provided in the WISC–V manual has been questioned by researchers (e.g., Canivez & Watkins, 2016). Finally, this study may prove useful for evaluating whether BSEM can offer greater insight into the nature of cognitive abilities and their relationship with existing tests of intelligence.

3. Method

3.1. Participants

Participants included a randomly selected sample ($N = 710$) of children between the ages of 6 and 16 years referred for clinical assessments through a large, outpatient pediatric psychology/neuropsychology clinic within a children's specialty hospital. Deidentified WISC–V ten primary subtest data were retrieved from the hospital's electronic medical records. Use of the data for inclusion in the study was approved by the hospital's Institutional Review Board.

Table 2 presents demographic characteristics of the clinical sample used in the analysis. As shown, the sample was primarily composed of White/Caucasian and Black/African American youth. The participants' ages ranged from 6.0 to 16.93 years and averaged 10.88 years ($SD = 2.79$ years). Table 3 presents the composition of the clinical sample demonstrating that three diagnostic groups (ADHD, 48.3 %; other nervous system disorders, 14.1 %; and anxiety, 10.1 %) comprised nearly three-fourths of the sample. Table A1[1] provides the descriptive statistics for the ten WISC–V subtests and corresponding index scores; Table A2 presents the covariance/correlation matrices; and Table A3 contains the Mplus code with discussion should a researcher wish to use this information to reproduce the analyses presented in this study.[2]

3.2. Instrument

The WISC–V contains 16 subtests, but its ten-subtest primary battery is typically administered in clinical practice (Benson et al., 2019). The scoring structure for the primary battery includes five indices: the Verbal Comprehension Index (VCI; Similarities and Vocabulary); Visual Spatial Index (VSI; Block Design and Visual Puzzles); Fluid Reasoning Index (FRI; Matrix Reasoning and Figure Weights); Working Memory Index (WMI; Digit Span and Picture Span); and Processing Speed Index (PSI; Coding and Symbol Search). Subtest scores have means of 10 with standard deviations of 3. Index scores have means of 100 with standard deviations of 15. Detailed descriptions of the WISC–V along with evidence preliminary reliability and validity evidence are available in the *WISC–V Technical and Interpretive Manual* (Wechsler, 2014b) and elsewhere (e.g., Kaufman et al., 2016; Sattler et al., 2016).

3.3. Procedure

BSEM (i.e., Bayes CFA) was used to investigate three different WISC–V models that have been featured within either the manual of the measurement instrument (e.g., the five factor higher-order model that is used by practitioners to score the instrument) or the extant literature (e.g., a four factor bifactor model that was found by independent research to be preferable to the publisher's presented five factor higher-order scoring structure). Mplus 8.4 (Muthén & Muthén, 1998–2017) was used for Bayesian estimation. Four different BSEM specifications were used to evaluate each of the models: (1) an analysis *without* cross-loadings or correlated residuals; (2) an analysis where all cross-loading are simultaneously estimated; (3) an analysis where all cross-loadings *and* correlated residuals for the subtests only are specified; and (4) an analysis where all cross-loadings and correlated residuals for the subtests and group factors are simultaneously estimated. A prior mean of 0 and variance of .01 was established a priori for cross-loadings based upon theoretical considerations. Given the interrelationship among cognitive ability subtest indicators this cross-loading range was posited to be appropriately large to detect meaningfully important cross-loadings, but not too large to cause issues with model convergence. A prior variance of .01 allows a cross-loading estimate range of −.20 to .20 to be recovered.

This study also conducted a sensitivity analysis for prior variances of .001, .005, .01, .02, .03, .04 and .05 respectively. A second sensitivity analysis was conducted investigating whether the parameter estimates were stable across iterations (I) in accord with Muthén and Asparouhov (2012). To be thorough, iterations of I, 2I, 4I, 8I, 10I, 20I and 25I, where $I = 10{,}000$, were evaluated using a prior of .01 across the cross-loadings only models. An Inverse-Wishart prior variance based on the procedure outlined in Asparouhov and Muthén (2015) was selected for specification of subtest residual prior variances (Asparouhov & Muthén, 2010) while that

[1] Tables denoted by "A" indicate supplementary materials, which can be found at the following link: https://osf.io/by28n/?view_only=fc350822960948ab8b8e35a2dd48b068

[2] Interested readers may also contact the lead author for any questions pertaining to the code used in this study.

Table 2
Demographic Characteristics of the Clinical Sample.

			Identified Sex	
Race/Ethnicity	*N*	Percent	Female	Male
White	365	51.4	120	245
Black	211	29.7	62	149
Hispanic	22	3.1	5	17
Multi-racial	67	9.4	23	44
Unknown/Other	45	6.3	20	25
Total	710		230	480
Percent		100.0	32.4 %	67.6 %

Table 3
Diagnostic Categories of the Clinical Sample.

ICD-10 Diagnosis	*N*	Percent
ADHD	343	48.3
Other nervous system disorders	100	14.1
Anxiety disorders	72	10.1
Adjustment disorder	36	5.1
Mood disorders	35	4.9
Epilepsy	26	3.7
Oncologic conditions	16	2.3
Disruptive behavior disorders	16	2.3
Other behavioral and emotional disorders	16	2.3
Other medical conditions	14	2.0
Learning/cognitive/speech disorders	14	2.0
Congenital abnormalities	11	1.5
Chromosomal abnormalities	6	0.9
Traumatic brain injury	4	0.5
Total	710	100.0

Note. ICD = international classification of diseases, tenth edition; ADHD = attention deficit/hyperactivity disorder

discussed by Muthén and Asparouhov (2012) was used for the group factor residuals to cohere with best practice. Two MCMC chains were used and iterations were established with the first half discarded as the burn-in phase. A model was determined to have attained convergence when the PSR stabilized on a value less than 1.10 and when there was a satisfactory Kolmogorov–Smirnov distribution (i.e., no discrepant posterior distributions in the different MCMC chains that led to model non-convergence; Muthén & Muthén, 1998–2017).

4. Results

The results of the iteration sensitivity analysis for all models at a prior of .01 for the cross-loadings only analysis is shown in Table 4. As indicated, the iteration sensitivity analysis produced model convergence and consistent DIC levels across the four and five factor higher-order models. Both models also produced consistent parameter estimates regardless of iteration selected. On the other hand, the four factor bifactor model produced model convergence at iteration levels between 40,000 and 100,000 but not below or above this level. The results of this sensitivity analysis also demonstrated a wider range in DIC and unstable parameter estimates for the bifactor model. A priors sensitivity analysis was also undertaken (see Table 5). As shown, the four factor bifactor model converged at a prior level of .005 and .01 but not at other levels. With the four and five factor higher-order models, a prior of .005 through .04 produced model convergence, and DIC stability. Although the 0.03 and 0.04 levels produced the lowest DIC for the four- and five-factor higher-order models respectively, a prior variance of 0.01 was deemed best for several reasons for all models: (1) it was the a priori established prior variance level based upon theoretical considerations and extant literature; (2) increasing the prior to a level higher than 0.01 did not materially alter the magnitude and patterning of loadings; and (3) a prior variance higher than 0.01 led to model nonconvergence for the correlated residuals analyses. Evidence of model convergence for the five-factor higher-order model with correlated residuals is shown in Figs. A1 and A2 in the online supplement (https://osf.io/by28n/?view_only=fc350822960948ab8b8e35a2dd48b068)where both the distribution and scatter plots suggested that the model fit the data well. In totality, the four- and five-factor higher-order models produced stable and consistent parameter estimates, whereas the four-factor bifactor model did not across both the iteration and prior variance sensitivity analyses. When the bifactor model was tested at different iterations and prior variance levels it displayed model instability and incoherence with the bifactor results from prior literature (see Table 6 for loading estimates demonstrating model instability).

Table 7 shows the results of the three models tested according to four specifications: (1) no cross loadings; (2) cross-loadings only; (3) cross-loadings plus correlated residuals (subtests); and (4) cross-loadings plus correlated residuals (subtests and group factors).

Table 4
Iteration Sensitivity (Prior Variance = 0.01) and Resulting DIC.

	Markov Chain Monte Carlo Iterations (MCMC)							DIC
Model	10 K	20 K	40K	80 K	100 K	200 K	250 K	Range
5 HO with Cross Loadings	NC	NC	NC	16,198	16,195	16,198	16,195	3
4 HO with Cross Loadings	16,223	16,222	16,215	16,218	16,218	16,219	16,219	8
4 BF with Cross Loadings	16,216	16,213	NC	NC	NC	16,155	16,155	61

Note. HO = Higher order, BF = Bifactor, NC = Non-convergence.
DIC = Deviance Information Criteria. Parameter estimates were stable across the 4 HO and 5 HO models.
4 BF parameter estimates unstable according to I, 2I, 4I, 8I, 10I, 20I and 25I;
4 and 5 HO have stable estimates across all iterations. Please see Table 6 for BF parameter (i.e., standardized loading) estimates.

Table 5
Sensitivity Analysis Priors (.001 to .05).

	Prior Variance						
Model	.001	.005	.01	.02	.03	.04	.05
5 HO with Cross Loadings							
PPP	.009	.293	.440	.547	.567	.535	NC
DIC	Reject	16,218	16,211	16,200	16,174	16,119	NC
4 HO with Cross Loadings							
PPP	.043	.212	.267	.343	.347	.343	.345
DIC	Reject	16,226	16,218	16,219	16,213	16,219	16,216
4 BF with Cross Loadings*							
PPP	NC	.46	.46	NC	NC	NC	NC
DIC	NC	16,175	16,155	NC	NC	NC	NC

Note. *BF model at .005 and .01 produced non-significant and negative group factor loadings for WM and PS.
For the .01 prior run, SI and VC also negatively and non-significantly loaded on VC for the BF model.
NC = Non-convergence, HO = Higher order, BF = Bifactor, PPP = Posterior predictive *p*-value, DIC=Deviance Information Criteria.

Table 6
Four Bifactor Group Factor Loadings.

Iterations	200 K		10 K	20 K
Prior Variance	.005	.01	.01	.01
Verbal				
Similarities	.45	−.44	**.44**	**.44**
Vocabulary	.45	−.44	**.44**	**.44**
Perceptual Reasoning				
Block Design	.46	**.43**	**.47**	**.43**
Visual Puzzles	.41	**.39**	**.43**	**.38**
Matrix Reasoning	.19	.21	.22	.19
Figure Weights	**.24**	.25	**.27**	.24
Working Memory				
Digit Span	−.26	−.28	.32	−.33
Picture Span	−.26	−.28	.32	−.33
Processing Speed				
Coding	**.55**	**.55**	**.55**	**.55**
Symbol Search	**.55**	**.55**	**.55**	**.55**

Note. ns = Non-significant ($p > .05$). **Bold** = significant parameter estimates, PRI=Perceptual Reasoning Index. Only 10 K, 20 K, 200 K and 250 K iterations converged and produced interpretable estimates.

There were several notable findings. Although the four-factor bifactor model (with cross-loadings) produced the lowest DIC (see Table 7), when the model was tested at different iterations and prior variance levels it displayed model instability and incoherence with prior literature (see Table 6). Specifically, the only group factor to consistently emerge was the Processing Speed factor. The other factors produced negatively loaded parameter estimates or did not significantly load their theoretically posited factors. The bifactor model also did not converge when correlated residuals were specified across all prior variance levels and iterations. Consequently, the bifactor model was deemed a generally poor fit with these data.

Also shown in Table 7, the five-factor higher-order model with cross-loadings only produced the second lowest DIC to that of the four-factor bifactor model. When correlated residuals were specified for the subtests, none produced a significant association across either the four- or five-factor higher-order models. When correlated residuals for subtests *and* latent group factors were specified, the five-factor higher-order model demonstrated a significant correlated residual between the Fluid Reasoning group factor and the Visual

Table 7
Model Comparison and Fit (Prior Variance = 0.01).

Models	PPP	DIC	95 % CrI Lower 2.5 %	95 % CrI Upper 2.5 %	pD	No. of Parameters
4 BF	.00	16,243	8.9	61	36.3	37
4 BF with Xloads* 200 K iterations	.46	16,155	−29.6	32.3	−17.7	67
4 BF with Xloads & Corr Residuals		**Model Rejected. PSR > 1.10**				
4 HO	.00	16,249	23.0	72.9	32.7	34
4 HO with Xloads 80 K iterations	.27	16,218	−20.8	39.4	36.4	64
4 HO with Xloads & Corr Resid (Subtests)	.59	16,225	−34.0	27.1	57.0	109
4 HO Xloads & Corr Resid (Subtests & Group factors)	.59	16,225	−34.6	27.6	56.7	115
5 HO	.00	16,275	37.0	96.4	35.1	35
5 HO with Xloads 80 K iterations	.44	16,198	−28.6	32.9	24.5	74
5 HO with Xloads & Corr. Residuals (Subtests)	.43	16,218	−28.3	33.3	43.7	119
5 HO Xloads & Corr Resid (Subtests & Group factors)	.57	16,220	−34.1	27.6	50.6	130

Note. PPP = Posterior predictive p value, DIC = Deviance information criteria, CrI = Credibility index, pD = Estimated number of parameters, BF = Bifactor, HO = Higher order, Xloads = Cross loadings, Corr resid = Correlated residuals. PSR = Potential scale reduction.

Spatial group factor (.44, $p = .034$). This suggests that Fluid Reasoning and Visual Spatial share sources of influence on the indicators that are unrelated to the factors (i.e., they contain unique information in common that is not accounted for by their respective factors). The specification of correlated residuals did not uncover any additional important relationships beyond the aforementioned correlated residual. Whether the cross-loadings only or the cross-loadings plus correlated residual analysis was evaluated, all specifications demonstrated no cross-loadings, similar magnitude and direction of primary loadings (see Table A4), and only one significant correlated residual estimate. The totality of these results suggested that, besides the correlated residual between FRI and VSI with the five-factor higher-order model, the WISC–V attained simple structure where the primary loadings significantly load on a single theoretically coherent factor.

Since there has been considerable debate and controversy surrounding whether the four-factor bifactor model discussed in the extant research literature is superior to the five-factor higher-order model presented in the WISC–V manual, and considering that BSEM is a relatively unknown methodology in school psychology (and psychology more broadly), both models were compared using Maximum Likelihood (ML) CFA (e.g., via the Satorra-Bentler correction due to multivariate non-normality of these data) to cross validate the BSEM finding that the five-factor higher-order model with the aforementioned correlated residual is the best fit (Table 8). Among the three models tested, the ML CFA results suggested that the five-factor higher-order model containing the correlated residual between the Fluid Reasoning and Visual Spatial factors produced the best fit with these data. This suggests that the BSEM analyses where correlated residuals were specified provided additional insight into the structure of the WISC–V ten subtest primary battery not previously uncovered in the extant frequentist literature. Importantly, it also provides some support with this sample for the scoring structure presented in the WISC–V manual in contrast with the conclusions from Canivez et al. (2017), Dombrowski et al. (2017) and others (e.g., Dombrowski et al., 2019; Watkins et al., 2018).

Table 9 presents the subtest loadings and variance estimates for the five-factor higher-order model, which was deemed the best model to represent the WISC–V ten subtest primary battery in the present sample. As shown in Table 9 and depicted in Fig. 1, the results display subtest loadings on theoretically consistent factors, no cross-loadings, and therefore the attainment of simple structure. The results are also consistent with extant structural validity research in cognitive assessment suggesting that primary interpretive emphasis should be placed upon the higher-order general factor as the *g* factor accounted for a higher percentage of variance (45 %) than that of the group factors which ranged from 11 to 14 %. While primary interpretive emphasis of the general factor (i.e., the FSIQ) should be regarded, this should not be misconstrued to suggest that the group factors (i.e., index level scores) should not receive *any* interpretive emphasis. Although the general factor variance accounts for much of the variance in the WISC-V there is still sufficient variance at the index level should psychologists wish to move to that level of interpretation when clinical situations demand the maximization of reliable explanatory variance in a survey-level assessment.

5. Discussion

Bayesian structural equation modeling was used to examine the structure of the WISC–V ten subtest primary battery with data obtained from a clinical sample. This analysis permitted a more nuanced and elaborate investigation than what could be obtained

Table 8
Frequentist Maximum Likelihood CFA Fit Statistics.

Model	S-B x^2	df	CFI	TLI	SRMR	RMSEA	RMSEA 90 % CI	BIC	AIC
5 Higher-Order	93.1	30	.984	.976	.027	.054	(.042–.067)	16,434	16,274
4 Bifactor	59.2	28	.992	.987	.020	.040	(.025–.054)	16,412	16,243
5 Higher-Order with correlated residual of FR with VS	**52.7**	**29**	**.994**	**.991**	**.019**	**.034**	**(.019–.048)**	**16,398**	**16,234**

Note. S-B = Satorra-Bentler, TLI = Tucker–Lewis Index, CFI = Comparative Fit Index, RMSEA = Root Mean Square Error of Approximation. AIC = Akaike's Information Criterion, BIC=Bayesian Information Criteria, FR = Fluid Reasoning, VS=Visual-Spatial.

Table 9
Five factor Higher Order Model with Cross-loadings and Correlated Residuals (0.01) 100 K Iterations.

	General		Verbal		VS		FR		WM		PS			
Subtest	b	s^2	b	s^2	b	s^2	b	s^2	b	s^2	b	s^2	h^2	u^2
Similarities	.66	.43	.77	.60									.71	.29
Vocabulary	.73	.53	.86	.73									.70	.30
Block Design	.76	.57			.84	.71							.60	.40
Visual Puzzles	.74	.55			.82	.67							.65	.35
Matrix Reasoning	.70	.48					.77	.59					.50	.50
Figure Weights	.68	.47					.75	.57					.79	.22
Digit Span	.71	.50							.80	.65			.64	.36
Picture Span	.58	.33							.66	.43			.74	.26
Coding	.50	.25									.66	.43	.47	.53
Symbol Search	.66	.44									.88	.77	.71	.29
Total Variance		.46		.13		.14		.12		.11		.12	.65	.35
Explained Common Variance		.43		.12		.13		.11		.10		.11		

Second-Order Loadings (Median)		Correlated Residual
Verbal	.85	Fluid Reasoning with Visual Spatial = .44
Visual-Spatial (VS)	.90	
Fluid Reasoning (FR)	.91	
Working Memory (WM)	.88	
Processing Speed (PS)	.76	

Note. b = standardized loading of subtest on factor (median), s^2 = variance, h^2 = communality, u^2 = uniqueness.
Small variance crossing loadings are in the range of .00 to .04, non-significant, and redacted for clarity.

using both traditional EFA and CFA (in combination) and produced some interesting findings. First, it demonstrated, perhaps more fully than any previous analysis, that the WISC–V is a well-constructed instrument that attained what Thurstone (1954) referred to as simple structure (i.e., subtest indicators load one and only one factor without any cross loadings). Relatedly, BSEM offered additional insight into the potential underlying complexity in theoretical structure of the WISC–V because of its unique analytical capabilities (e.g., simultaneous estimation of all cross-loadings and correlated residual terms). Until recently, this was unavailable to researchers investigating the relationship of intelligence theory to available tests of intelligence. Because of this analytical technology it could be useful to intelligence researchers who wish to better understand intelligence theory and intelligence test factor structure. With respect to the WISC–V, BSEM confirmed that there were no instances of covariation among constructs except for the finding of a correlated residual between the Fluid Reasoning and Visual Spatial group factors, which has been posited in previous CFA research on the broader 16 subtest total battery configuration in the normative sample (e.g., Reynolds & Keith, 2017). Whereas Reynolds and Keith (2017) arrived at this conclusion using multiple model specifications this finding was accomplished parsimoniously with a single run. Regarding the interpretation of this finding, it may well be conceptualized as an intermediate factor between the general factor and the group factors or it could be thought of as a latent construct that explains something between two latent variables apart from the general factor. Alternatively, it simply could be statistical noise where there is a finding between two latent variables and such finding is unanticipated or perhaps meaningless. Put simply, there is no definitive psychological explanation as to what a correlated residual represents (particularly at the factor-level) beyond theoretical conjecture by the researcher. Whereas relations between a core test and a recall measure makes intuitive sense the prior explanations remain speculative without additional modeling to verify these hypothetical structural complexities. Additional research on the topic of correlated residual use in BSEM is necessary.

Second, the results of BSEM indicated that the four-factor bifactor model appeared to be unstable with this clinical sample. The model either does not consistently converge or produces parameter estimates (i.e., loadings) that varied depending upon iteration and prior variance selection. This is an interesting finding and something that Dombrowski, McGill, and Morgan (2021) also observed in a very small number of replications when undertaking a Monte Carlo simulation of the WISC–V normative data and the normative data for other cognitive ability instruments. However, this finding is inconsistent with a recent large body of frequentist factor analytic research that generally supports the bifactor model as offering the preferred structure for the WISC–V and other cognitive ability instruments such as the WJ–IV Cognitive, the DAS–II, and the KABC–II (c.f. Dombrowski, McGill, & Morgan, 2021; McGill et al., 2018). For instance, numerous researchers across multiple WISC–V studies have concluded that a four bifactor conceptualization of the WISC–V provided the best fit with the data (e.g., Canivez et al., 2017; Pauls & Daseking, 2021). These studies noted that the WISC–V structure is reminiscent of the four-factor structure of the WISC–IV that contained Verbal Comprehension, Working Memory, Processing Speed and Perceptual Reasoning (i.e., fusion of Visual-Spatial and Fluid Reasoning into a complexly determined Perceptual Reasoning factor). The rationale for the disparity between BSEM and frequentist factor analytic procedures pertaining to the bifactor model needs further study given the implications for the clinical interpretation (Rodriguez et al., 2016a). Although the purpose of this study is not to adjudicate this issue in particular, a potential explanation for these findings could relate to the nature of the underlying assumptions of the bifactor model. The bifactor model may have limitations for evaluating the structure of cognitive attributes (Reynolds & Keith, 2013) when it is statistically under identified and in need of constraining. The inclusion of numerous specifications,

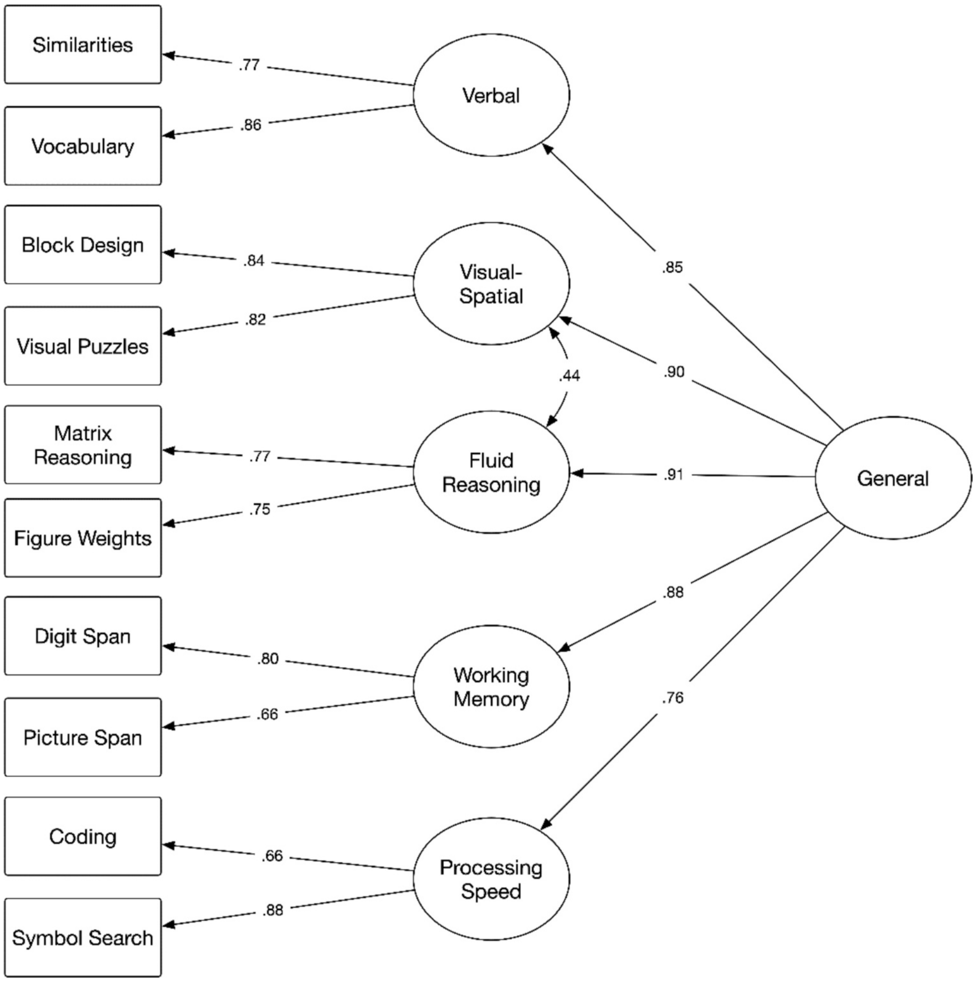


Fig. 1. Five Factor Higher Order BSEM Validation Model for the WISC–V Primary Subtests with a Clinical Sample.
Note. g = general intelligence. All standardized loading estimates (median) are statistically significant. Residual terms are omitted for clarity.

even small variance priors, could cause the model to be over-identified and fail to properly converge (Dombrowski et al., 2019; Zhang et al., 2021). A replication of the results of this study using a larger number of subtests (e.g., the normative sample's 16 subtest primary and secondary battery) would be worthwhile.

Third, when this study used the results of a Bayesian analysis to guide a subsequent maximum likelihood CFA analysis comparing the five-factor higher-order model to the four-factor bifactor model, the five-factor higher-order model containing the specification of a correlation in the residuals between FRI and VSI produced the best fit with the data. This is an intriguing finding and one that was not uncovered previously in the literature for the ten subtest primary battery.[3] It suggests that BSEM was able to locate the 'best' model fit without need for multiple modifications often undertaken in cognitive ability research (e.g., Beaujean, 2016; Reynolds & Keith, 2017) that might give the appearance of hypothesizing after the results are known (i.e., HARKing; Kerr, 1998), a practice that should be delimited when attempting to replicate or reproduce the structure of an existing measure as it can be antithetical to the scientific practice of falsification (Popper, 1962) and prone to confirmation bias (Brown, 2015; Kahneman, 2011).

Fourth, Stromeyer et al. (2015) have criticized the use of cross-loadings and correlated residuals contending that their specification simply adds statistical noise, clutters a model with nonsensical information, and detracts from simple structure. Notwithstanding the implication that this same criticism can be levied (erroneously) against an entire class of factor analysis (e.g., EFA) with a long standing and deep history, the results of this study suggest the opposite to the conclusion posed by Stromeyer et al. given that the specification of cross-loadings and correlated residuals clarified, not obscured, the structure of the WISC–V as an instrument that is free of cross-loadings and correlated residuals save for one with theoretical meaning. BSEM suggested that either the four- or the five-factor higher-order models are parsimonious and fit these data well though the five-factor higher-order model had the lowest DIC.

[3] It should be noted that it was disclosed in the manual (Wechsler, 2014b) that this parameter improved model fit for the 16 subtests total battery configuration, but its retention was rejected.

Fifth, BSEM offers contemporary intelligence researchers a technology unavailable to prior generations (e.g., Carroll, Spearman, Horn, Cattell) who used prevailing methodology (e.g., EFA, CFA) accessible to them at the time to model the latent structure of cognitive abilities. Accordingly, BSEM may provide researchers in intelligence theory a methodological approach that can examine additional interrelationships among abilities not captured by frequentist structural equation modeling methods, addressing a concern noted by McGrew et al. (2023) and Kovacs and Conway (2016). As a hybrid of EFA and CFA, BSEM may be useful in attenuating, but not fully eliminating, the practice of specification searches endemic to structural equation modeling which may capitalize on chance (Meehl, 1993). For instance, BSEM can circumscribe the sequential practice of modifying parameters post hoc, permitting a concurrent examination of all specifications (e.g., cross-loadings and correlated error terms). This is one of the powerful features of BSEM that could be useful for the modeling of human cognitive abilities, which are known to have many overlapping and interrelated components that often require elaborate post hoc modeling to uncover. The results from a BSEM analysis may even be useful in informing scoring structures for various instruments which assume simple structure and may not take into consideration instances where the underlying structure of a test may be more theoretically complex.

Sixth, the practical implications of this study suggest that the WISC–V ten-subtest primary battery readily attains simple structure and might be adequately scored as intended by the test publisher (i.e., interpret the FSIQ and five index level scores). This contrasts with the majority of independent structural validity literature, which suggested that a four-factor bifactor model akin that that of the WISC-IV (Wechsler, 2003) is preferred whereby the Fluid Reasoning and Visual Spatial subtests coalesce to form a combined FRI/VSI (formerly the Perceptual Reasoning) factor. Although the present BSEM results tacitly support the underlying scoring structure of the ten WISC–V primary subtests from a theoretical perspective, it is important to evaluate sources of variance (e.g., explained total and common variance; see Table 9) when determining the adequacy of measurement for the interpretation of scores (Reise et al., 2023; Rodriguez et al., 2016a; Rodriguez et al., 2016b; Sellbom & Tellegen, 2019). The variance ascribed to the general factor (represented by the FSIQ) and the five groups factors suggests that much of the interpretive emphasis should be placed upon the FSIQ score but that the five index scores have sufficiently meaningful variance for subsequent interpretation beyond the FSIQ.

Finally, Muthén and Asparouhov (2012) discussed the utility of BSEM for scale development, where researchers can use it in a stepwise fashion to modify and improve upon the instrument. There is an additional potentially important use. Given its ability to freely estimate more parameters than conventional frequentist methods, BSEM may help to resolve the replication crisis surrounding the theoretical structure of cognitive ability measures generally and the WISC-V specifically. Since the WISC–V was first published, a variety of research has emerged suggesting a different factor structure for the WISC–V (e.g., a four-factor bifactor model; Dombrowski et al., 2019) than that posited by the test publisher. Considering the popularity of the instrument and its centrality in high-stakes clinical decisions (e.g., identification of intellectual disability and specific learning disability), this is a problem. When multiple studies fail to replicate or reproduce results, then this undermines confidence in an instrument's structure and how that instrument should be scored and interpreted.

6. Limitations

As with any method, BSEM is not without limitations. While it is important to understand the structure of tests with clinical samples—most youth who receive such measures are part of these samples—it may not be possible to fully generalize these results to a normative population or in focal clinical situations that do not cohere with the sample in question. In particular, the current study's sample is predominantly comprised of youth with a primary diagnosis of Attention-Deficit/Hyperactivity Disorder which could result in altered structures due to the underlying cognitive deficits associated with the disorder (Becker et al., 2024).

Most saliently, BSEM requires a higher level of statistical coding sophistication and access to raw data. It also is a technique that requires further study, critique, and cross validation with modeling approaches of the frequentist variety given its limited use to this point. One of the advantages of BSEM—the specification of priors—also poses a potential limitation. When attempting to replicate or reproduce the structure of an existing instrument, or when attempting to create a new assessment instrument, the selection of a prior should be established empirically to avoid specification searches where researchers chose a prior solely because the researcher prefers one model (e.g., higher-order vs bifactor) over another. This haphazard approach to model fitting should be eschewed (Dombrowski et al., 2022; Meehl, 1993). As a result, use of BSEM can attenuate, but does not fully resolve, the practice of post hoc specification that has been staunchly criticized in structural validity research since the inception of modern computational techniques (Horn, 1989).

Further, additional research needs to be conducted regarding use of correlated residuals in BSEM. There have only been two prior studies using BSEM on commercial ability measures (e.g., Dombrowski et al., 2018; Golay et al., 2013) and a select few investigating psychology, health, and management (De Bondt & Van Petegem, 2015; Fong & Ho, 2013, 2014; Stromeyer et al., 2015; Zyphur & Oswald, 2015). Some researchers raised concerns about the use of correlated residuals claiming that their use does nothing more than add statistical noise (Stromeyer et al., 2015). However, Muthén and Asparouhov (2012) and Asparouhov and Muthén (2015) contend that if correlated residuals are used appropriately then their specification will enhance the understanding of an instrument's underlying structure. The present study demonstrated that their use does not appear to add statistical noise and obscure model fit; instead, it appeared to provide a degree of structural and theoretical clarity regarding the WISC–V increasing confidence that the instrument may measure what the test publisher claims it to measure as well as cohere with models produced by independent researchers (e.g., Reynolds & Keith, 2017). In fact, the specification of correlated residuals led to the uncovering of a relationship between VSI and FRI, which, when subsequently modeled in the five factor higher-order models using maximum likelihood CFA, produced a marginally better fit than any other tested model. Whether this reflects capitalization on chance or whether BSEM represents an improvement over the combined use of EFA and CFA requires further evaluation. It does appear, however, to be an intriguing alternative to modeling methods of the frequentist variety and should benefit from increased attention particularly when used to study the structure of contemporary intelligence tests.

7. Conclusion

As indicated in this study the use of BSEM offered greater insight into the factor structure and theory of one of the world's most commonly used assessment instruments, the WISC–V, by showing that the instrument attains simple structure and may reflect the theoretical five group factors suggested by the test publisher. While some may think this is an obvious conclusion, it is not. The factor structure of the WISC–V has been vigorously debated in the assessment literature with most of that literature not only questioning the structure posited by the test publisher, but also offering alternative theoretical/factor structures, many of which have yet to be substantively replicated (Dombrowski, in press). This lack of convergence in the empirical literature is concerning. The resulting factor structure of an instrument assists with determining how that instrument should be scored and interpreted (Brunner et al., 2012; Dombrowski, McGill, Canivez, et al., 2021). When multiple studies fail to converge upon the same structure, then this undermines confidence in the instrument's posited scoring approach and suggests a replication problem (Dombrowski & McGill, 2024). Use of BSEM holds promise for helping to streamline the recovery of plausible structural models for an assessment instrument by obviating the need for post hoc tweaking in conventional frequentist methods when there is additional complexity in an underlying model. These results illustrate that traditional approaches for recovering posited model complexity may simply produce rival models for an instrument that obscures its true underlying structure and that are unlikely to replicate in subsequent research. Though BSEM requires additional statistical understanding and coding sophistication, the extra effort may prove worthwhile particularly when a variety of potential rival theoretical and interpretive models emerge for an instrument in the assessment literature.

Informed Consent/Patient Consent: No patients were used in the writing of this manuscript so informed

Consent is not required.

Code

Furnished in the online appendix of this study to permit reproduction.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

All procedures were conducted in accordance with the ethical standards in publishing.

CRediT authorship contribution statement

Stefan C. Dombrowski: Conceptualization, Methodology, Software, Formal Analysis, Data Curation, Writing – original draft, Writing – review & editing. **Ryan J. McGill:** Writing – original draft, Writing – review & editing. **Gary L. Canivez:** Writing – original draft, Writing – review & editing. **Marley W. Watkins:** Writing – original draft. **Alison E. Pritchard:** Resources, Writing – review & editing, Writing – original draft. **Lisa A. Jacobson:** Resources, Writing – review & editing, Writing – original draft.

Declaration of competing interest

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jsp.2025.101432.

Data availability

Covariance/correlation matrices are available to permit reproduction of this study.

References

Asparouhov, T., & Muthén, B. (2010). Bayesian analysis using Mplus: Technical implementation. Retrieved from https://www.statmodel.com/download/Bayes3.pdf.

Asparouhov, T., Muthén, B., & Morin, A. J. S. (2015). Bayesian structural equation modeling with cross-loadings and residual covariances: Comments on Stromeyer et al. (2015). *Journal of Management, 41*(6), 1561–1577. https://doi.org/10.1177/0149206315591.

Beaujean, A. A. (2016). Reproducing the Wechsler Intelligence Scale for Children-Fifth Edition: Factor model results. *Journal of Psychoeducational Assessment, 34*(4), 404–408. https://doi.org/10.1177/0734282916642679.

Becker, A. B. C., Maurer, J., Daseking, M., & Pauls, F. (2024). Measurement invariance of the WISC-V across a clinical sample of children and adolescents with ADHD and a matched control group. *Journal of Intelligence, 12*(1), 6. https://doi.org/10.3390/jintelligence12010006.

Benson, N. F., Floyd, R. G., Kranzler, J. H., Eckert, T. L., Fefer, S. A., & Morgan, G. B. (2019). Test use and assessment practice of school psychologists in the United States: Findings from the 2017 national survey. *Journal of School Psychology, 72*, 29–48. https://doi.org/10.1016/j.jsp.2018.12.004.

Brown, T. A. (2015). *Confirmatory factor analysis for the applied researcher*. Guilford Press.

Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *Journal of Personality, 80*(4), 796–846. https://doi.org/10.1111/j.1467-6494.2011.00749.x.

Cain, M. K., & Zhang, Z. (2019). Fit for a Bayesian: An Evaluation of PPP and DIC for Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 26*(1), 39–50. https://doi.org/10.1080/10705511.2018.1490648.

Canivez, G. L. (2016). Bifactor modeling in construct validation of multifactored tests: Implications for understanding multidimensional constructs and test interpretation. In K. Schweizer, & C. DiStefano (Eds.), *Principles and methods of test construction: Standards and recent advancements* (pp. 247–271). Gottingen, Germany: Hogrefe.

Canivez, G. L., & Kush, J. C. (2013). WISC–IV and WAIS–IV structural validity: Alternate methods, alternate results. Commentary on Weiss et al. (2013a) and Weiss et al. (2013b). Journal of Psychoeducational Assessment, 31(2), 157–169. doi: https://doi.org/10.1177/0734282913478036.

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. In, S. E. Kaufman, & D. L. Coalson (Eds.), *Intelligent testing with the WISC-V* (pp. 683–702). Wiley.

Canivez, G. L., Watkins, M. W., & Dombrowski, S. C. (2017). Structural validity of the Wechsler Intelligence Scale for Children-Fifth Edition: Confirmatory factor analyses with the 16 primary and secondary subtests. *Psychological Assessment, 29*, 458–472. https://doi.org/10.1037/pas0000358.

Carroll, J. B. (1985). Exploratory factor analysis: A tutorial. In D. K. Detterman (Ed.), *Current topics in human intelligence, Vol. 1: Research methodology (pp. 25–58)*. Ablex.

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

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician, 46*(3), 167–174. https://doi.org/10.2307/2685208.

De Bondt, N., & Van Petegem, P. (2015). Psychometric evaluation of the Overexcitability Questionnaire-Two applying Bayesian structural equation modeling (BSEM) and multiple-group BSEM-based alignment with approximate measurement invariance. *Frontiers in Psychology, 6*, 1–17. https://doi.org/10.3389/fpsyg.2015.01963.

Dombrowski, S.C. (in press). Contributing to the Reproducibility Crisis in Psychology: The Role of Statistical Software Choice on Factor Analysis. Journal of School Psychology.

Dombrowski, S. C., Beaujean, A. A., Schneider, J. W., McGill, R. J., & Benson, N. (2019). Using exploratory bifactor analysis to understand the latent structure of multidimensional psychological measures: An applied example featuring the WISC–V. *Structural Equation Modeling: A Multidisciplinary Journal, 26*(6), 847–860. https://doi.org/10.1080/10705511.2019.1622421.

Dombrowski, S. C., Canivez, G. L., & Watkins, M. W. (2017). Factor structure of the 10 WISC–V primary subtests across four standardization age groups. *Contemporary School Psychology, 22*, 90–104. https://doi.org/10.1007/s40688-017-0125-2.

Dombrowski, S. C., Golay, P., McGill, R. J., & Canivez, G. L. (2018). Investigating the theoretical structure of the DAS-II core battery at school age using Bayesian structural equation modeling. *Psychology in the Schools, 55*(2), 190–207. https://doi.org/10.1002/pits.22096.

Dombrowski, S. C., & McGill, R. J. (2024). Clinical Assessment in School Psychology: Impervious to Scientific Reform? Canadian Journal of School Psychology, 0(0). doi: https://doi.org/10.1177/08295735231224052.

Dombrowski, S. C., McGill, R. J., Canivez, G. L., Watkins, M. W., & Beaujean, A. A. (2021). Factor analysis and variance partitioning in intelligence research: Clarifying misconceptions. *Journal of Psychoeducational Assessment, 39*(1), 28–38. https://doi.org/10.1177/0734282920961952.

Dombrowski, S. C., McGill, R. J., Farmer, R. L., Kranzler, J. H., & Canivez, G. L. (2022). Beyond the rhetoric of evidence-based assessment: A framework for critical thinking in clinical practice. *School Psychology Review, 51*(6), 771–784. https://doi.org/10.1080/2372966X.2021.1960126.

Dombrowski, S. C., McGill, R. J., & Morgan, G. W. (2021). Monte Carlo modeling of contemporary intelligence test (IQ) factor structure: Implications for IQ assessment, interpretation and theory. *Assessment, 28*(3), 977–993. https://doi.org/10.1177/1073191119869828.

Edwards, M. C. (2010). A Markov chain Monte Carlo approach to confirmatory item factor analysis. *Psychometrika, 75*, 474–497. https://doi.org/10.1007/s11336-010-9161-9.

Fong, T. C., & Ho, R. T. (2013). Factor analyses of the Hospital Anxiety and Depression Scale: A Bayesian structural equation modeling approach. *Quality of Life Research, 22*, 2857–2863.

Fong, T. C., & Ho, R. T. (2014). Testing gender invariance of the Hospital Anxiety and Depression Scale using the classical approach and Bayesian approach. *Quality of Life Research, 23*(10), 1421–1426. https://doi.org/10.1007/s11136-013-0594-3.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* ((2nd ed.).). Chapman and Hall/CRC.

Gelman, A., Meng, X.-L., & Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6*, 733–807. https://www.jstor.org/stable/24306036.

Gelman, A., & Rubin, D. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science, 7*(3), 457–511. https://www.jstor.org/stable/2246093.

Golay, P., Reverte, I., Rossier, J., Favez, N., & Lecerf, T. (2013). Further insights on the French WISC–IV factor structure through Bayesian structural equation modeling. *Psychological Assessment, 25*(2), 496–508. https://doi.org/10.1037/a0030676.

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

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika, 82*(4), 711–732. https://doi.org/10.2307/2337340.

Hoofs, H., van de Schoot, R., Jansen, N. W. H., & Kant, I. (2018). Evaluating model fit in Bayesian confirmatory factor analysis with large samples: Simulation study introducing the BRMSEA. *Educational and Psychological Measurement, 78*(4), 537–568. https://doi.org/10.1177/0013164417709314.

Horn, J. (1989). Models of intelligence. In R. L. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 29–75). University of Illinois Press.

Kahneman, D. (2011). *Thinking, fast and slow*. Straus and Giroux: Farrar.

Kaplan, D., & Depaoli, S. (2013). Bayesian statistical methods. In T. Little (Ed.), *1. The Oxford handbook of quantitative methods* (pp. 407–437). Oxford University Press.

Kaufman, A. S., Raiford, S. E., & Coalson, D. L. (2016). *Intelligent testing with the WISC–V*. Wiley.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4.

Kovacs, K., & Conway, A. R. A. (2016). Process overlap theory: A unified account of the general factor of intelligence. *Psychological Inquiry, 27*(3), 151–177. https://doi.org/10.1080/1047840X.2016.1153946.

Link, W. A., & Eaton, M. J. (2012). On thinning of chains in MCMC. *Methods in Ecology and Evolution, 3*(1), 112–115. https://doi.org/10.1111/j.2041-210X.2011.00131.x.

Loehlin, J. C., & Beaujean, A. A. (2017). *Latent variable models: An introduction to factor, path, and structural equation analysis* ((5th ed.).). Taylor & Francis.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structures analysis: The problem of capitalizing on chance. *Psychological Bulletin, 111*(3), 490–504. https://doi.org/10.1037//0033-2909.111.3.490.

Marsh, H. W., Muthén, B., Asparouhov, A., Ldtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling, 16*(3), 439–476. https://doi.org/10.1080/10705510903008220.

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. https://doi.org/10.1016/j.jsp.2018.10.007.

McGrew, K. S., Schneider, W. J., Decker, S. L., & Bulut, O. (2023). A psychometric network analysis of CHC intelligence measures: Implications for research, theory, and interpretation of broad CHC scores "beyond *g*". *Journal of Intelligence, 11*(1), 19. https://doi.org/10.3390/jintelligence11010019.

Meehl, P. E. (1993). Four queries about factor reality. *History and Philosophy of Psychology Bulletin, 5*(2), 4–5.

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods, 17*(3), 313–335. https://doi.org/10.1037/a0026802.

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* ((8th ed.).). Muthén & Muthén.
Pauls, F., & Daseking, M. (2021). Revisiting the factor structure of the German WISC-V for clinical interpretability: An exploratory and confirmatory approach on the 10 primary subtests. *Frontiers in Psychology, 12*, Article 710929. https://doi.org/10.3389/fpsyg.2021.710929.
Popper, K. (1962). *Conjectures and refutations: The growth of scientific knowledge* ((2nd ed.).). Routledge.
R Core Team. (2023). *R: A language and environment for statistical computing*. Available online at: R Foundation for Statistical Computing. https://www.Rproject.org.
Reise, S. P., Mansolf, M., & Haviland, M. G. (2023). Bifactor measurement models. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (2nd ed., pp. 329–348). New York: Guilford Press.
Reynolds, M. R., & Keith, T. Z. (2013). Measurement and statistical issues in child assessment research. In D. H. Saklofske, C. R. Reynolds, & V. L. Schwean (Eds.), *The Oxford handbook of child psychological assessment* (pp. 48–83). 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. https://doi.org/10.1016/j.intell.2017.02.005.
Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016a). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*(2), 137–150. https://doi.org/10.1037/met0000045.
Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016b). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment, 98*(3), 223–237. https://doi.org/10.1080/00223891.2015.1089249.
Sattler, J. M., Dumont, R., & Coalson, D. L. (2016). *Assessment of Children: WISC–V and WPPSI-IV*. Sattler Publisher.
van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods, 22*(2), 217–239. https://doi.org/10.1037/met0000100.
Sellbom, M., & Tellegen, A. (2019). Factor analysis in psychological assessment research: Common pitfalls and recommendations. *Psychological Assessment, 31*(12), 1428–1441. https://doi.org/10.1037/pas0000623.
Stone, J. V. (2013). *Bayes' rule: A tutorial introduction to Bayesian analysis*. Sebtel Press.
Stromeyer, W. R., Miller, J. W., Sriramachandramurthy, R., & DeMartino, R. (2015). The prowess and pitfalls of Bayesian structural equation modeling: Important considerations for management research. *Journal of Management, 41*(2), 491–520. https://doi.org/10.1177/0149206314551962.
Thompson, B. (2004). Exploratory and confirmatory factor analysis: Understanding concepts and applications. *American Psychological Association.* https://doi.org/10.1037/10694-000.
Thurstone, L. L. (1954). An analytical method for simple structure. *Psychometrika, 19*, 173–182. https://doi.org/10.1007/BF02289182.
Vehtari, A., & Gelman, A. J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*, 1413–1432. https://doi.org/10.1007/s11222-016-9696-4.
Watkins, M. W. (2018). Exploratory factor analysis: A guide to best practice. *Journal of Black Psychology, 44*(3), 219–246. https://doi.org/10.1177/0095798418771807.
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 & Educational Psychology, 6*(4), 252–265. https://doi.org/10.1080/21683603.2017.1342580.
Wechsler, D. (2003). *Wechsler Intelligence Scale for Children-Fourth Edition*. Psychological Corporation.
Wechsler, D. (2014a). *Wechsler Intelligence Scale for Children-Fifth Edition*. NCS Pearson.
Wechsler, D. (2014b). *Wechsler Intelligence Scale for Children-Fifth Edition technical and interpretive manual*. NCS Pearson.
Zhang, B., Sun, T., Cao, M., & Drasgow, F. (2021). Using bifactor models to examine the predictive validity of hierarchical constructs: Pros, cons, and solutions. *Organizational Research Methods, 24*(3), 530–571. https://doi.org/10.1177/1094428120915522.
Zyphur, M. J., & Oswald, F. L. (2015). Bayesian estimation and inference: A user's guide. *Journal of Management, 41*(2), 390–420. https://doi.org/10.1177/0149206313501200.

Online Supplement with Addendum

Conjectures and Refutations in Cognitive Ability Structural Validity Research: Insights from Bayesian Structural Equation Modeling

By S. C. Dombrowski et al.

Table A1

Descriptive Statistics for WISC-V Subtest and Index Scores

Score	*N*	Mean	SD	Skewness	Kurtosis
Block Design	710	8.6	3.5	+0.23	-0.02
Similarities	710	9.2	3.4	-0.01	-0.17
Matrix Reasoning	710	9.1	3.4	-0.07	-0.30
Digit Span	710	8.0	3.1	+0.22	+0.26
Coding	710	7.3	3.3	+0.06	-0.36
Vocabulary	710	8.9	3.6	+0.06	-0.58
Figure Weights	710	9.6	3.2	+0.05	-0.29
Visual Puzzles	710	9.7	3.4	+0.02	-0.34
Picture Span	710	8.6	3.3	+0.10	-0.30
Symbol Search	710	8.1	3.3	+0.00	-0.01
VCI	702	94.8	18.0	-0.03	-0.25
VSI	703	95.3	17.9	0.17	-0.03
FRI	702	96.2	17.3	-0.03	-0.44
WMI	703	90.1	16.2	0.14	-0.16
PSI	703	87.0	17.1	-0.10	-0.02
FSIQ	670	90.8	17.7	+0.04	-0.27

Note: VCI=Verbal Comprehension Index; VSI=Visual Spatial Index; FRI=Fluid Reasoning Index; WMI=Working Memory Index; PSI=Processing Speed Index; FSIQ=Full Scale Intelligence Quotient.

Table A2

Covariance and Correlations Matrices

Covariance Matrix

	BD	SI	MR	DS	CD	VO	FW	VP	PS	SS
BD	11.922									
SI	6.28	11.724								
MR	7.2	5.929	11.634							
DS	5.231	6.023	5.536	9.596						
CD	4.663	3.974	4.421	4.425	10.784					
VO	7.035	9.336	6.442	6.408	4.702	12.857				
FW	7.177	6.137	6.785	5.001	3.782	6.673	10.17			
VP	8.647	6.573	7.254	5.427	4.625	7.188	7.062	11.381		
PS	5.007	4.916	5.115	5.783	4.12	6.025	4.683	5.279	10.689	
SS	5.323	4.468	4.734	4.502	6.704	5.094	3.903	5.272	4.214	10.841

Correlation Matrix

	BD	SI	MR	DS	CD	VO	FW	VP	PS	SS
BD	1									
SI	0.531	1								
MR	0.611	0.508	1							
DS	0.489	0.568	0.524	1						
CD	0.411	0.353	0.395	0.435	1					
VO	0.568	0.76	0.527	0.577	0.399	1				
FW	0.652	0.562	0.624	0.506	0.361	0.584	1			
VP	0.742	0.569	0.63	0.519	0.417	0.594	0.656	1		
PS	0.444	0.439	0.459	0.571	0.384	0.514	0.449	0.479	1	
SS	0.468	0.396	0.422	0.441	0.62	0.431	0.372	0.475	0.391	1

Table A3

Mplus Input Code

```
Title:    BSEM WISC-5 Clinical Sample 10 Subtest
          !Model 5 Higher Order OR 5 Bifactor Model (VCI, VZI FRI, WMI, PSI)
data:
file is "WISCV.txt";
VARIABLE:
NAMES ARE bd si mr ds cd vo fw vp ps ss;
USEVARIABLES ARE bd si mr ds cd vo fw vp ps ss ;
define:
         standardize bd si mr ds cd vo fw vp ps ss  ;
analysis:
estimator = bayes;
!estimator = mlm;
proc = 2;
fbiter = 100000;
chain = 2
stvalues = ml;
Kolmogorov = 100;

MODEL:
! 5  Higher Order
vci by si vo ;
vzi by bd vp;
fri by mr fw;
wmi by ds ps ;
psi by cd ss;
g by vci vzi fri wmi psi ;

! For bifactor model, cross loads and correlated residuals code remains the same but replace with the followig code:
! 5 Bifactor

!vci by si* vo (1);
!vci@1;

!vzi by bd* vp (2);
!vzi@1;

!fri by mr* fw (3);
!fri@1;

!wmi by ds* ps (4);
!wmi@1;

!psi by cd* ss (5);
!psi@1;

!g by si* vo bd vp mr fw ds ps cd ss ;
!g@1;

!vci with vzi-g@0;
!vzi with fri-g@0;
!fri with wmi-g@0;
!wmi with psi-g@0;
!psi with g@0;


!cross-loadings:
!   Saying *0 gives a zero mean start value for the parameter.
!   The prior with mean zero and small variance (e.g., .01) is then applied during the computations.

vci by bd*0 vp*0 mr*0 fw*0 ds*0 ps*0 cd*0 ss*0(xload1-xload8);

vzi by si*0 vo*0 mr*0 fw*0 ds*0 ps*0 cd*0 ss*0 (xload9-xload16);

fri by si*0 vo*0 bd*0 vp*0 ds*0 ps*0 cd*0 ss*0 (xload17-xload24);

wmi by si*0 vo*0 bd*0 vp*0 mr*0 fw*0 cd*0 ss*0 (xload25-xload32);

psi by si*0 vo*0 bd*0 vp*0 mr*0 fw*0 ds*0 ps*0 (xload33-xload40);


!correlated error terms

bd-ss(p1-p10);

!bd-ss are the residual variances of the subtests.
!The IW prior is assigned for the residual variance using b1-b10 as described below.
```

```
!These are the subtest covariances lower-triangular elements taken row-wise

     bd-ss with bd-ss (c1-c45);
     vci-psi(a1-a5);
     vci-psi with vci-psi(b1-b10);

!b1=error covariance; a1=error variance; IW defined below and assigned as noted above
!b1 a1
!b2 b1 a2
!b3 b1 b2 a3
!b4 b1 b2 b3 a4

     model priors:

!Create the residual variance-covariance matrix (by default, residuals are independent, thus not correlated).
!This will allow to estimate them because otherwise they wouldn't be part of the model.
!There will be 45 covariances and 10 residual variances in total. Forumla: n(n-1)/2 for # of covar.
!Name the variances from b1-b10 and the covariances from c1-c45


!This is the approach described in Muthen & Asparouhov (2012).

!16 = 10 PARAM + 6 !Assign an inverse-wishart prior to the residual variances
!(following Muthen and Asparouhov instructionsn see page 14-15)
!df=# of parameters +6; where IW (I, df)

!p1-p10~IW(1,16);

!####################################################################################################################
!Steps to conduct corr residuals based on Asparouhov et al (2015). See Asparouhov et al (2015) for technical description in appendix

!Steps to conduct correlated residuals.
!1) Run xloads only analysis.
!2) Obtain residual var from each indicator.
!3) The forumla is IW(residual var*df, df) or (.282*100, 100).!Df is obtained from Asparouhov et al. (2015) where
!df is chosen based upon sample size such that N=10,000 then df=1,000. N=500 the df =100.
!It takes an additional step and specification as noted above and reference to their response to Stromeyer et al (2015) to obatin df.
!In this example since N=710, this is about ~500 so d =100 for this formula.

          p1~IW(28.2, 100);
          p2~IW(27.3, 100);
          p3~IW(43.5, 100);
          p4~IW(31.2, 100);
          p5~IW(46.9, 100);
          p6~IW(20.7, 100);
          p7~IW(37.9, 100);
          p8~IW(26.2, 100);
          p9~IW(51.8, 100);
          p10~IW(26.9, 100);

!Assigns IW prior to latent group residual variances.
          a1-a5~IW(1,11);

!Assign another kind of inverse-wishart prior to the subtest covariances (Asparouhove &Muthen, 2015)
          c1-c45~IW(0,100);

!Assigns an IW to latent group factor covariances based on Muthen & Asparouhov (2012)
          b1-b10~IW(0, 11);

!Assigns prior variance to the xloads

          xload1-xload40~N(0, .01); !Assigns prior variance of .01 to the xloads


          output:
          tech1 tech8 stdy svalues;

          plot:
          type = plot2;
```

Table A4

Five Factor Higher Order Model Loading Estimates

	No priors	Xloads Only	Xloads & Corr Resid (Subtests)	Xloads & Corr Resid (Subtests & Group Factors)
Verbal				
SI	.85	.81	.65	.77
VO	.89	.78	.89	.86
Visual Spatial				
BD	.46	.94	.89	.85
VP	.41	.89	.94	.82
Fluid Reasoning				
MR	.78	.77	.80	.75
FW	.80	.84	.74	.75
Working Memory				
DS	.80	.74	.74	.79
PS	.71	.53	.53	.66
Processing Speed				
CD	.75	.72	.82	.66
SS	.82	.80	.71	.88
General				
VCI	.84	.78	.77	.85
VSI	.92	.95	.96	.91
FRI	.97	.98	.99	.90
WMI	.86	.75	.78	.88
PSI	.69	.66	.66	.76

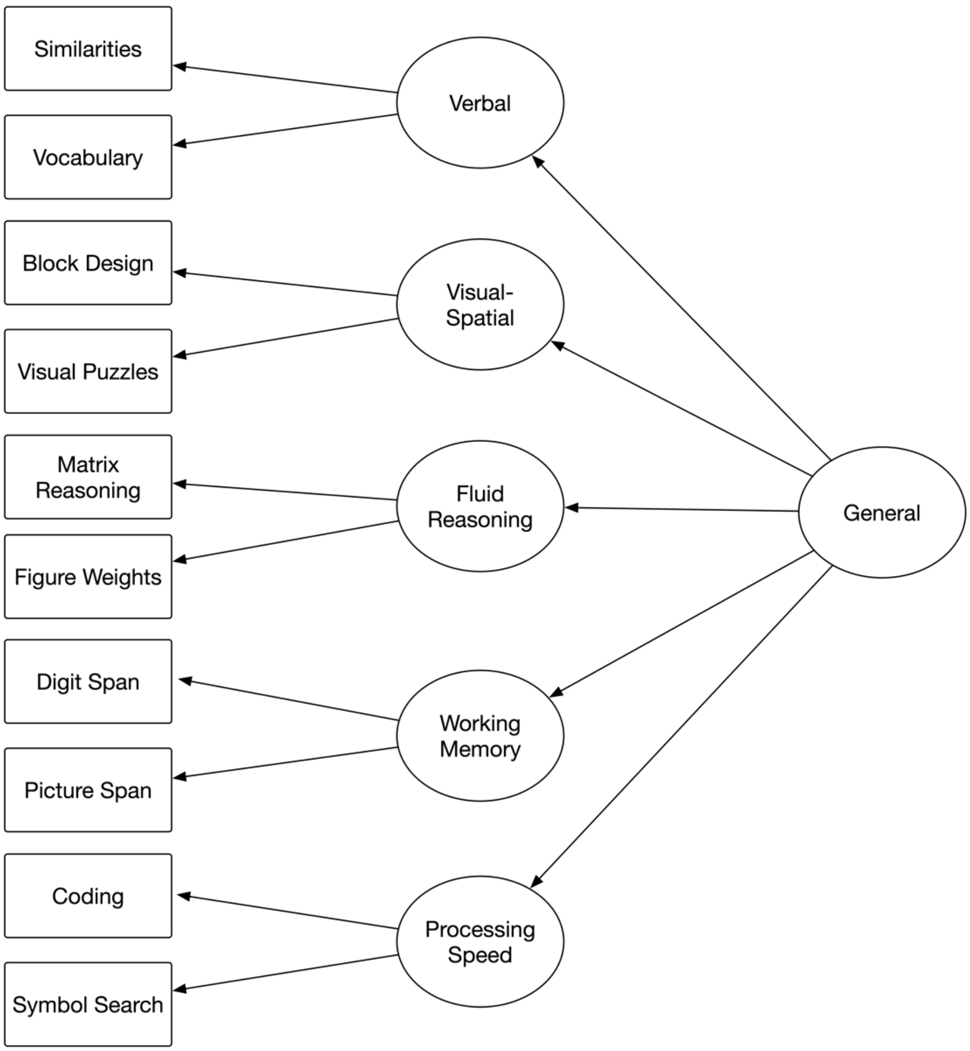
Note. Xloads=Cross-loadings. Corr resid=Correlated residual. Correlated residual between FRI and VSI (.432) significant at $p<.05$

Addendum

Higher-Order vs. Bifactor Models: Distinctions with a *Slight* Difference

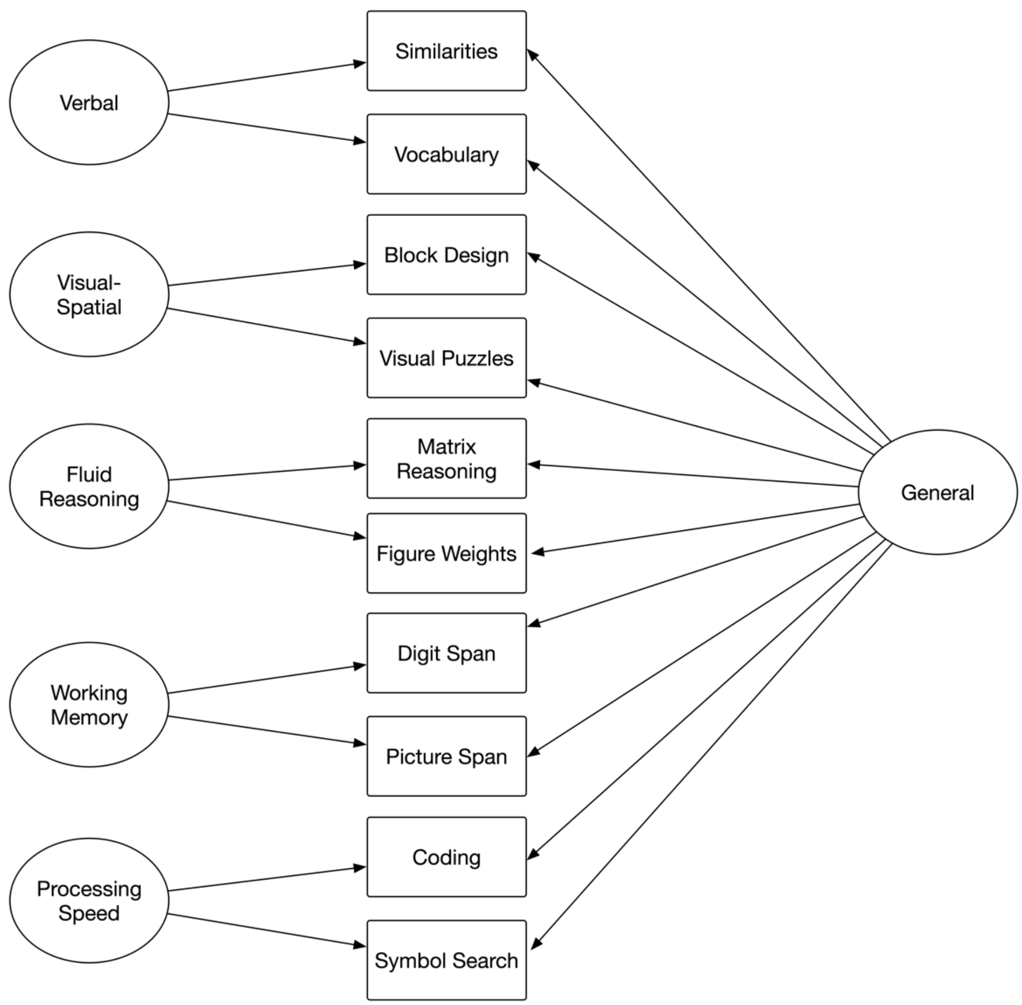
When modelling the factor structure of IQ tests, researchers generally consider two different structural models. The first is the higher-order (second-order) model where the general (second-order) factor is fully mediated by first-order group factors in influencing subtests indicators (measured variables) below the group factors. This model is depicted in Figure 1 below using the WISC-V ten subtest primary battery.

Figure 1. Higher Order Model for the WISC-V



The second is the bifactor model where the general factor and group factors simultaneously have direct influences on individual subtest indicators (measured variables). This is depicted in Figure 2 below using the WISC-V ten subtest battery.

Figure 2. Bifactor Model for the WISC-V



While both models acknowledge the existence of the general factor (*g*) on intelligence tests, the higher-order model conceptualizes the general factor as superordinate to the group factors and subtests with first-order group factors between the general factor and the individual subtests (measured variables). In other words, the general factor has no direct influence on the individual subtests as its influence is fully mediated by the first-order group factors. Conversely, the bifactor model conceptualizes the general factor as a breadth factor and assumes that *g* and the group factors have simultaneous direct effects on the subtests. For a more technical discussion of this topic please see Beaujean (2015), Canivez (2016), and Gignac (2008). From a practical, interpretive perspective, the bifactor model provides partitioned variance to determine the relative influence of the general factor versus group factors which has implications for their use in interpretive methods which stress primary interpretation of group factor indices (e.g., PSW). This may be accomplished with a higher order model (as per Keith & Reynolds, 2018) but the merits of variance partitioning using the HO model[1] have not been thoroughly investigated so this practice should be considered experimental until further verification ensues.

From a theoretical perspective the higher order model's conceptualization of *g* is akin to looking at the shadow of a person standing next to a street light versus looking directly at the person (bifactor) to ascertain how tall they are. The higher-order model was explicated by Thurstone (1947) and is considered a bottom-up (American model) where group factors are prioritized (see Beaujean & Benson, 2019 for a discussion). Spearman (1904) was among the first to discuss *g* where he posited a two-factor theory for the construct. With Spearman's model (i.e., the British approach) interpretation of the general factor is prioritized and group/specific factors are regarded as largely a statistical nuisance. However, in some of his later writings, Spearman regarded group/specific factors with greater importance. In fact, one of Spearman's post-doctoral fellows, Karl Holzinger, elaborated on two-factor theory culminating in the development of what later became known as the bifactor model (Holzinger & Swineford, 1937).

In 1957, Schmid and Leiman created their orthogonalization procedure. The SL procedure represents an elegant transformation of the higher-order model and is considered an approximate bifactor model (Reise, 2012). The bifactor conceptualization of intelligence essentially lay dormant until 1993 when John Carroll created his magnum opus, *Human Cognitive Abilities,* where he re-analyzed approximately 457 datasets going back to the 1920s. This creation ostensibly served as a bulwark against those who disavowed the importance and even existence of *g*. The creation of the SL transformation also raised awareness of the bifactor model as a tenable model for contemporary intelligence tests, but could only be produced secondarily from standard EFA results.

In 2011, Jennrich and Bentler created a true exploratory version of the bifactor model via analytic rotation (for further application and simulation of its use on real world data see Dombrowski et al., 2021). Prior to that time researchers who wanted an exploratory bifactor modeling approach had to utilize the Schmid-Leiman (1957) procedure. At present, both the higher-order model and the bifactor model may be used to investigate the structure of tests of cognitive ability though application of the former to other psychological measures is controversial as the theoretical justification for a general factor of say personality is less well-developed (Bonifay et al., 2017). Despite debate (see Decker et al. [2021] and a response via Dombrowski et al. [2021]) about which model reflects the true reality of the structure of tests of cognitive abilities and the nature of intelligence, there have been few studies that have empirically tested this issue (see Dombrowski, McGill, & Morgan [2021] for a Monte Carlo simulation of the structure of all major intelligence tests) so the issue remains yet unresolved.

References

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

Beaujean, A. A., & Benson, N. F. (2019). The one and the many: Enduring legacies of Spearman and Thurstone on intelligence test score interpretation. *Applied Measurement in Education, 32*(3), 198-215. https://doi.org/10.1080/08957347.2019.1619560

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*(1), 184-186. https://doi.org/10.1177/2167702616657069

Canivez, G. L. (2016). Bifactor modeling in construct validation of multifactored tests: Implications for understanding multidimensional constructs and test interpretation. In K.

[1] It is acknowledged that the Schmid-Leiman (1957) orthogonalization procedure is predicated upon the higher-order model transforming it into an approximate bifactor model where variance is apportioned to the general factor and group factors. This is different than Keith and Reynold's (2018) approach which apportions variance and preserves higher-order structure.

Schweizer & C. DiStefano (Eds.), *Principles and methods of test construction: Standards and recent advances* (pp. 247-271). Hogrefe.

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

Decker, S. L., Bridges, R. M., Luedke, J. C., & Eason, M. J. (2021). Dimensional evaluation of cognitive measures: Methodological confounds and theoretical concerns. *Journal of Psychoeducational Assessment, 39*(1), 3-27. https://doi.org/10.1177/073428292094087

Dombrowski, S. C., Beaujean, A. A., Schneider, J. W. & McGill, R. J. & Benson, N. (2019). Using exploratory bifactor analysis to understand the latent structure of multidimensional psychological measures: An applied example featuring the WISC-V. *Structural Equation Modeling: A Multidisciplinary Journal, 26*(6), 847-860. https://doi.org/10.1080/10705511.2019.1622421

Dombrowski, S. C., McGill, R. J., Canivez, G. L., Watkins, M. W., & A. A. Beaujean (2021). Factor analysis and variance partitioning in intelligence research: Clarifying misconceptions. *Journal of Psychoeducational Assessment, 39*(1), 28-28. https://doi.org/10.1177/0734282920961952

Dombrowski, S. C., McGill, R. J. & Morgan, G. W. (2021). Monte Carlo modeling of contemporary intelligence test (IQ) factor structure: Implications for IQ assessment, interpretation and theory. *Assessment, 28*(3), 977-993. https://doi.org/10.1177/1073191119869828

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

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika,* 2, 42-54. https://doi.org/10.1007/BF02287965

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

Keith, T. Z., & Reynolds, M. R. (2018). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (4th ed., pp. 853-900). Guilford Press.

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

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral*

Research, 47(5), 667-696. https://doi.org/10.1080/00273171.2012.715555

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

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

Thurstone, L. L. (1931). Multiple factor analysis. *Psychological Review, 38*(5), 406-427. https://doi.org/10.1037/h0069792

Figure A1

Posterior Predictive Checking Scatterplot

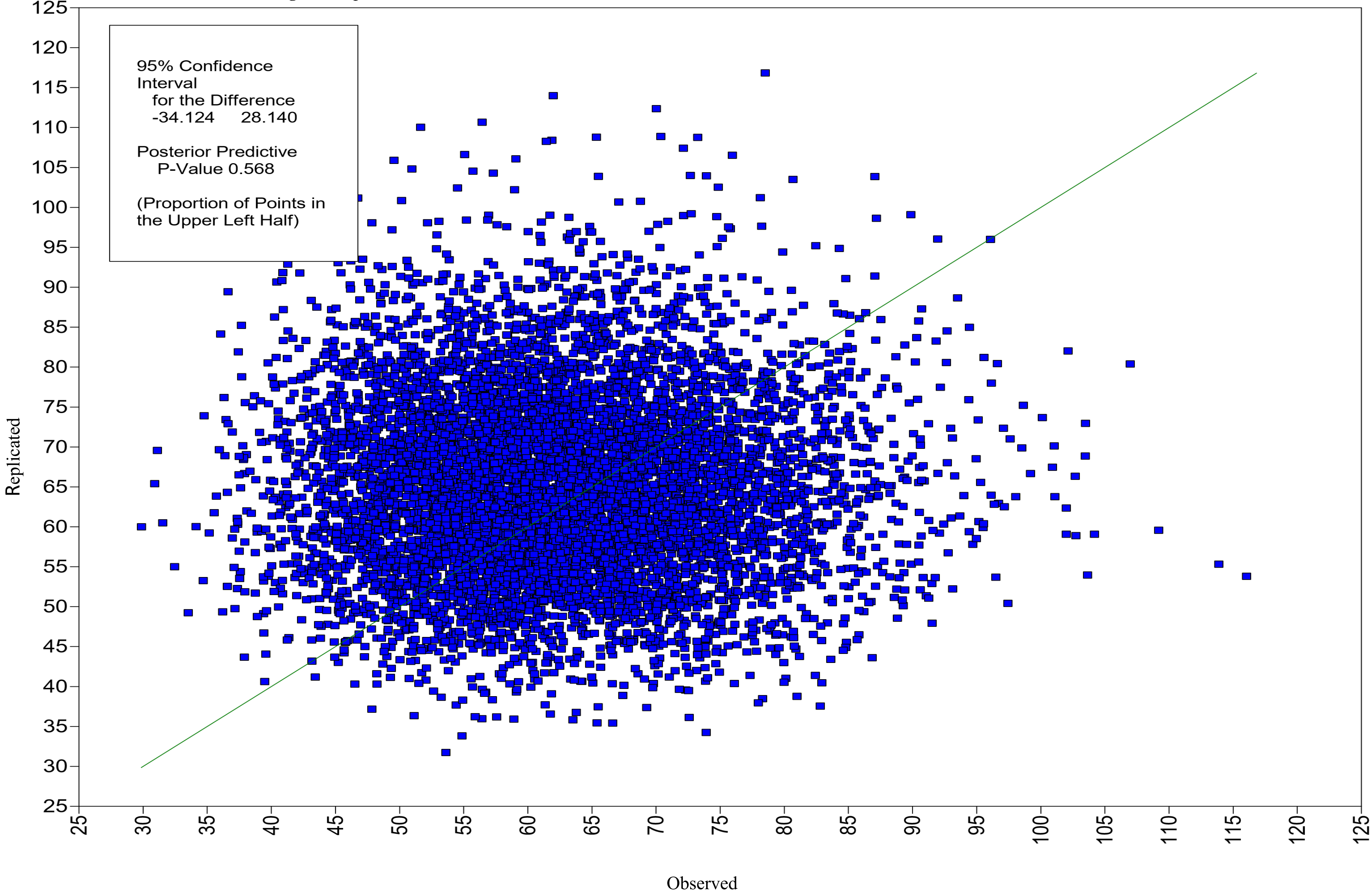


Figure A2

Posterior Predictive Checking Distribution Plot