INTERNATIONAL JOURNAL OF SCHOOL & EDUCATIONAL PSYCHOLOGY
2018, VOL. 6, NO. 4, 252–265
https://doi.org/10.1080/21683603.2017.1342580

ARTICLE

Reliability and factorial validity of the Canadian Wechsler Intelligence Scale for Children–Fifth Edition

Marley W. Watkins[a], Stefan C. Dombrowski[b], and Gary L. Canivez[c]

[a]Department of Educational Psychology, Baylor University, Waco, Texas, USA; [b]School Psychology Program, Rider University, Lawrenceville, New Jersey, USA; [c]Department of Psychology, Eastern Illinois University, Charleston, Illinois, USA

ABSTRACT

The reliability and factorial validity of the Wechsler Intelligence Scale for Children–Fifth Edition: Canadian (WISC-V^{CDN}) was investigated. The higher-order model preferred by Wechsler (2014b) contained five group factors but lacked discriminant validity. An alternative bifactor model with four group factors and one general factor, akin to the traditional Wechsler model, exhibited the best global fit. The general factor accounted for 33.8% of the total variance and 67.6% of the common variance, but none of the group factors accounted for substantial portions of variance. All together, the general and group factors accounted for 50% of the total variance. Omega reliability coefficients demonstrated that reliable variance of WISC-V^{CDN} factor index scores was primarily due to the general factor, not the group factors. It was concluded that the cumulative weight of reliability and validity evidence suggests that psychologists should focus their interpretive efforts at the general factor level and exercise extreme caution when using group factor scores to make decisions about individuals.

KEYWORDS
WISC-V; CFA; bifactor; intelligence; validity; reliability

The administration and interpretation of standardized psychological tests to assess cognitive functioning is a fundamental responsibility of psychologists (Australian Psychological Society [APS], 2009; Canadian Psychological Association [CPA], 2007; Evers et al., 2012; Hsu, Huang, & Cheng, 2009; Kranzler, 2016; Kranzler, Benson, & Floyd, 2016). Because "incompetent action is unethical per se" (CPA, 2000, p. 15), psychological assessment requires that psychologists have "a thorough understanding of statistics and psychometrics" sufficient to comprehend "the technical merits of selected instruments in terms of such characteristics as validity, reliability, standardization and test construction" (CPA, 2007, p. 8). Similar demands for understanding and applying psychometric evidence to test selection and interpretation have been articulated in other professional standards (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014; APS, 2009; British Psychological Society [BPS], 2007; Evers et al., 2013; International Test Commission [ITC], 2001; Krishnamurthy et al., 2004).

Psychometric competence is important when evaluating new assessment instruments (CPA, 2007). However, it is likely that "new" cognitive instruments encountered by psychologists will be revisions of existing instruments. In that case, Beaujean (2015a) recommended that a revision be treated as a new test because scores from the two instruments cannot be assumed to be directly comparable without supporting evidence. For example, subtest patterns and psychometric characteristics have been found to differ across Wechsler versions (Benson, Beaujean, & Taub, 2015; Strauss, Spreen, & Hunter, 2000).

Canadian WISC-V

A salient "new" test, the Wechsler Intelligence Scale for Children–Fifth Edition: Canadian (WISC-V^{CDN}; Wechsler, 2014a), was recently published. According to Beaujean (2015a), its scores cannot be assumed to be identical to those of its predecessor and its psychometric merits must be independently evaluated by prospective users (AERA, APA, & NCME, 2014; CPA, 2007). This seems especially appropriate because the WISC-V^{CDN} was a major revision involving the addition of three new subtests and two new factor indices, deletion of two subtests, and changes to the contents and instructions of all remaining subtests (Wechsler, 2014b).

Considerable psychometric information on the WISC-V^{CDN} was provided in its manual (Wechsler, 2014b). However, sole reliance on the opinion of test

CONTACT Marley W. Watkins ✉ Marley_Watkins@baylor.edu Department of Educational Psychology, Baylor University, Waco, TX 76798-7301, USA.

authors and publishers "is akin to relying solely on the opinions provided by pharmaceutical companies to make decisions on whether to take their medication. While their information can be valuable, these individuals . . . have a conflict of interest" (Beaujean, 2015a, p. 53). Additional information about the WISC-V^{CDN} can be obtained from independent test reviews. For example, the WISC-V^{CDN} was recently reviewed and found to maintain "many of the strong practical and psychometric qualities of its predecessors," but the size of its norming sample was criticized as insufficient (Cormier, Kennedy, & Aquilina, 2016, p. 332).

Although this review of the WISC-V^{CDN} characterized its psychometric properties as strong, it did not critically analyze the psychometric evidence presented by Wechsler (2014b) to support its claims of reliability and validity of the WISC-V^{CDN}. Nor was that psychometric evidence critically evaluated in a review of the U.S. version of the WISC-V (Na & Burns, 2016). Specifically, psychometric evidence presented by Wechsler (2014b, 2014c) was uncritically reported with no evaluation of the methods used to estimate reliability nor the degree to which the scoring structure of the WISC-V matched the theoretical structure of the underlying constructs it purports to measure—that is, evidence to support its factorial or structural validity (Messick, 1995). These omissions are puzzling, given that methodological and practical problems with similar reliability and validity methods have previously been reported (Canivez & Kush, 2013; Canivez, Watkins, & Dombrowski, 2016, 2017; Dombrowski, Canivez, Watkins, & Beaujean, 2015).

Validity

Regarding factorial validity, Wechsler (2014b) proposed a higher-order structure for the WISC-V^{CDN} with an overarching general intelligence (*g*) factor being loaded by five general factors which, in turn, were loaded by 16 primary and secondary subtests. This structure is illustrated in Figure 1 and was obtained via confirmatory factor analysis (CFA). However, "CFA studies based upon weak theoretical perspectives, lack of testing alternative theoretical views, or insufficient evidence may not offer adequate support of construct validity" (DiStefano & Hess, 2005, p. 225).

Guided by best practices in CFA (Bowen & Guo, 2012; Brown, 2015; DiStefano & Hess, 2005; Kline, 2016; MacCallum & Austin, 2000; McDonald & Ho, 2002; Widaman, 2012), there are six notable concerns regarding the CFA methods reported by Wechsler (2014b). First, not all plausible WISC-V^{CDN} models were tested by Wechsler (2014b). Inclusion of alternative conceptualizations of test structure is essential to provide convincing support for one model over another (Brown, 2015). Failure to include alternative models makes researchers susceptible to confirmation bias (MacCallum & Austin, 2000). That is, prone to seek or interpret "evidence in ways that are partial to existing beliefs" (Nickerson, 1998). For example, Wechsler (2014b, p. 53) stated that the preferred five-factor model "is supported by strong model fit indicators and consistently high factor loadings." However, Figure 1 clearly shows standardized factor loadings that are not generally considered to be high (e.g., .19–.34; DiStefano & Hess, 2005; Widaman, 2012).

The potential for confirmation bias may have been heightened by the nature of the models selected for evaluation of the WISC-V^{CDN}. According to Wechsler (2014b), CFA was used "to confirm whether the final factor model specified in the U.S. WISC-V could be applied to the Canadian WISC-V" (p. 48). Thus, only those models previously selected by the publisher of the U.S. WISC-V were included. A further caution concerns an intrinsic limitation of CFA. Namely, CFA can demonstrate that a model is

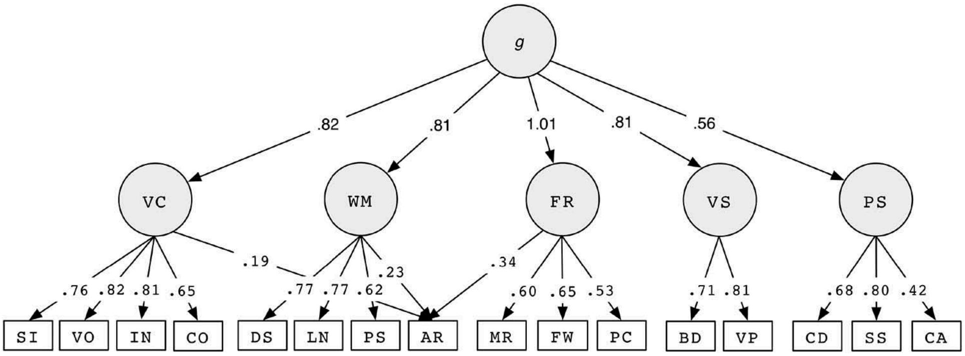

Figure 1. Standardized structure of the WISC-V^{CDN} proposed by Wechsler (2014b).

SI = Similarities, VO = Vocabulary, IN = Information, CO = Comprehension, DS = Digit Span, LN = Letter-Number Sequencing, PS = Picture Span, AR = Arithmetic, MR = Matrix Reasoning, FW = Figure Weights, PC = Picture Concepts, BD = Block Design, VP = Visual Puzzles, CD = Coding, SS = Symbol Search, CA = Cancellation, VC = Verbal Comprehension factor, WM = Working Memory factor, FR = Fluid Reasoning factor, VS = Visual Spatial factor, PS = Processing Speed factor, and *g* = General Intelligence.

consistent with the data but "it does not *confirm* the veracity of the researcher's model" (Kline, 2016, p. 21).

There is a variety of factorial models that can represent the structure of cognitive abilities. Those that include a general intelligence factor can be roughly separated into higher-order versus bifactor (sometimes called nested or direct hierarchical) models (Beaujean, 2015b). Figure 2 illustrates those two types of models. The higher-order model conceptualizes *g* as a superordinate factor having a direct effect on several group factors but only an indirect effect on the measured variables. Thus, the relationship of general intelligence to the measured variables is fully mediated by the group factors. In contrast, the bifactor model conceptualizes *g* as a breadth factor having direct effects on the measured variables, as do the group factors. Wechsler (2014b) only evaluated higher-order factor models for the WISC-V^{CDN}. Carroll's (1993) cognitive model, which is best represented through a bifactor model (Beaujean, 2015b), was incorporated into the Cattell-Horn-Carroll theory (CHC; Schneider & McGrew, 2012) that was considered in development of the WISC-V^{CDN} (Wechsler, 2014b). Additionally, bifactor models have been found to be good representations of the structure of other tests of cognitive ability (Brunner, Nagy, & Wilhelm, 2012; Canivez, 2014; Cucina & Howardson, 2017; Dombrowski et al., 2015; Gignac & Watkins, 2013; Golay, Reverte, Rossier, Favez, & Lecerf, 2013; Watkins & Beaujean, 2014). Finally, the bifactor model is a viable candidate for measures that have demonstrated good fit to a second-order model (Reise, 2012) and offers advantages for subsequent investigations of reliability and external validity (Benson, Kranzler, & Floyd, 2016; Brown, 2015; Canivez, 2016; Chen, Hayes, Carver, Laurenceau, & Zhang, 2012; Reise, Moore, & Haviland, 2010). Therefore, the omission of bifactor models from the CFA of Wechsler (2014b) calls into question the scoring structure of the WISC-V^{CDN} (Wasserman & Bracken, 2013).

Second, the method used to scale the latent variables in CFA models was not disclosed by Wechsler (2014b). In CFA, the distribution of latent variables (i.e., factors) and disturbance terms (i.e., factor variances) requires that some components be fixed in order to scale the latent variables and allow the model to be statistically

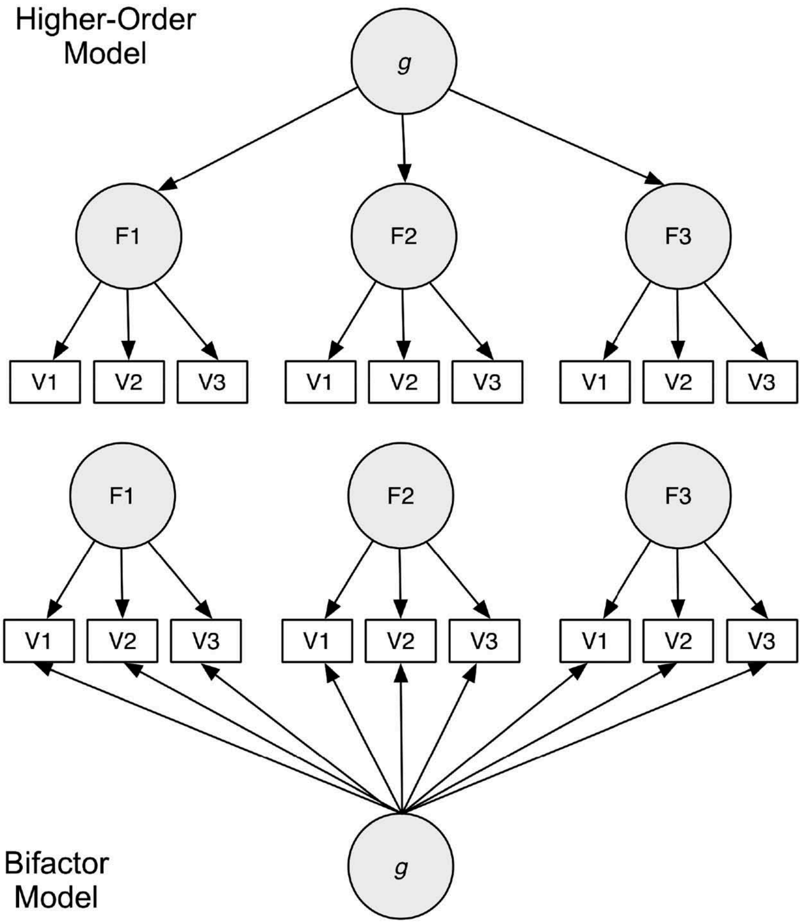

Figure 2. Conceptual illustration of higher-order versus bifactor models.

identified. The choice of scaling method can affect unstandardized parameters and may "yield different conclusions regarding the statistical significance of freely estimated parameters" (Brown, 2015, p. 133). Typically, either a reference indicator loading or a latent variable variance is fixed. Beaujean (2016) replicated the results reported in Wechsler (2014c) and deduced that an effects-coding method (Little, Slegers, & Card, 2006) was probably used with the U.S. WISC-V. All three methods should produce the same global model fit measures (i.e., chi-square, degrees of freedom, approximate fit indices) but Beaujean (2016) demonstrated that a modification of the effects-coding method was used in U.S. WISC-V analyses (2014c) that caused degrees of freedom to be understated with cascading consequences for fit statistics that rely on degrees of freedom for their computation. Beaujean (2016) concluded that this modified effects-coding method "should be employed with caution because it could produce multiple problems" (p. 406). Given that the analyses reported by Wechsler (2014b) for the Canadian WISC-V mirror those found in Wechsler (2014c) for the U.S. WISC-V, it appears that this modified effects-coding method was also employed in the WISC-V^{CDN} CFA. The consequences of using a nonstandard scaling method in the WISC-V^{CDN} analyses are unknown but should be explored through replication using conventional scaling methods.

Third, the method of estimating parameters in the CFA models tested by Wechsler (2014b) was nonstandard. "The choice of estimation method becomes essential because it will affect evaluation of model fit and parameter estimates" (Lei & Wu, 2012). Maximum likelihood (ML) estimation is typically used with multivariate continuous data and either weighted least squares (WLS) or robust ML with nonnormal data (Brown, 2015). Wechsler (2014b) reported that WLS estimation via SAS software was employed in CFA of the WISC-V^{CDN}. In its default version, the SAS WLS estimator is an asymptotic distribution-free estimator that is biased with sample sizes as small as those found in the WISC-V^{CDN} standardization sample (Lei & Wu, 2012). When considering estimators for continuous data, Hoyle (2000, p. 478) cautioned that "the use of an estimator other than maximum likelihood requires explicit justification" and Brown (2015, p. 346) concluded that "WLS is not recommended." Given these caveats, use of WLS is perplexing and represents a departure from the use of ML estimation typically observed in CFA of intelligence tests. Its effect is unknown but should be investigated.

Fourth, the preferred five-factor model abandoned the parsimony of simple structure (Thurstone, 1947) by allowing multiple cross-loadings of the Arithmetic subtest (see Figure 1). Simple structure facilitates interpretation because each variable loads onto only one factor. From a broader perspective, the parsimony of simple structure honors "the purpose of science [which] is to uncover the relatively simple deep structure principles or causes that underlie the apparent complexity observed at the surface structure level" (Le, Schmidt, Harter, & Lauver, 2010, p. 112). Following this concept, cross-loadings that are not both statistically and practically significant (i.e., > .30) might best be constrained to zero (Stromeyer, Miller, Sriramachandramurthy, & DeMartino, 2015). In fact, simple structure is implied by the scoring structure of the WISC-V^{CDN} where composite scores are created from unit-weighted sums of separate subtest scores. Thus, the preferred five-factor model is discrepant from the scoring structure of the WISC-V^{CDN}. Additionally, this complex structure can create identification problems because it deviates from an independent clusters structure (McDonald & Ho, 2002), especially when the primary index scales are formed by only two subtests.

Fifth, selection of the preferred five-factor model was primarily based on chi-square difference tests of nested models. Wechsler (2014b) acknowledged the sensitivity of the chi-square test of exact fit to trivial differences when analyzing large samples, but nevertheless used chi-square difference tests of nested models to identify the preferred five-factor model. However, large samples also make the chi-square difference test sensitive to trivial differences. In fact, Millsap (2007) admonished that "ignoring the global chi-square tests while at the same time conducting and interpreting chi-square difference tests between nested models should be prohibited as nonsensical." Additionally, there was no control of Type I error levels for these multiple statistical tests (Shaffer, 1995). For example, Wechsler (2014b) reported 9 chi-square difference tests at the .05 alpha level, which with a simple Bonferroni correction would suggest that each test should be set at .006 to maintain a study-wide error rate at the .05 level. Hence, the differences in global fit relied on by Wechsler (2014b) to select preferred models might reflect only trivial differences between models.

A somewhat related sixth concern is that global model fit, by itself, is an inadequate measure of model veracity (Bowen & Guo, 2012; Brown, 2015; DiStefano & Hess, 2005; Kline, 2016; MacCallum & Austin, 2000; Widaman, 2012). Even with good global fit, relationships among variables might be weak, parameter estimates might not be statistically significant, the latent variables might not account for meaningful variance in the indicators, or parameter values might be out of range (e.g., factor loadings that exceed 1.00, negative variance estimates, etc.). Even if not out of range, latent variables with nonsignificant variances "are not useful measures because they do not capture meaningful differences among individuals" (Bowen & Guo, 2012, p. 147). Wechsler (2014b) did not report the statistical significance of parameter estimates but a review of Figure 1, the publisher-preferred structural model for the WISC-V^{CDN}, reveals a standardized path coefficient of 1.01 between the higher-order general intelligence factor and the first-order Fluid Reasoning (FR) factor as well as a negative variance estimate (–.01) for the FR factor. This suggests that the *g* and FR factors were empirically redundant (Le et al., 2010), which constitutes a major threat to discriminant validity (Brown, 2015) and signals that the WISC-V^{CDN} may have been overfactored (Frazier & Youngstrom, 2007). Further, to search for better global model fit by testing mulitple variations of well-fitting models as done by Wechsler (2014b) may capitalize on sampling error and lead to final models that are not generalizable (Myung, 2000). Relatedly, Wechsler (2014b) failed to report the proportions of variance accounted for by general and group factors, nor the communality of measured variables. These statistics speak to the relationships of measured and latent variables and may be important for accurate interpretation of common factors (Brown, 2015; MacCallum & Austin, 2000).

Reliability

Concerns regarding reliability of WISC-V^{CDN} scores are interrelated with concerns about the validity of its scores because reliability is necessary but not sufficient for validity (Geisinger, 2013). Different estimates of reliability assess different sources of scoring inconsistency. For example, test–retest coefficients tap the consistency of responses across time while internal consistency coefficients meter the consistency of responses across test content. Nevertheless, all reliability estimates are a property of the scores on a test for a specific group of examinees (Geisinger, 2013). That is, estimates might differ from sample to sample. This distinction is important because Wechsler (2014b) did not compute test–retest reliability estimates for the WISC-V^{CDN} standardization sample. Rather, stability coefficients from the U.S. standardization of the WISC-V were reported for the WISC-V^{CDN}. This substitution was especially meaningful for the speeded subtests that comprise one of the composite index scores. Consequently, the reliability of those scores among Canadian children is unknown.

Wechsler (2014b) reported split-half coefficients computed from the standardization sample as indices of reliability for nonspeeded WISC-V^{CDN} subtests. The resulting coefficients ranged from .83 to .94, whereas composite index score coefficients based on those subtests ranged from .91 to .96. Wechsler (2014b) characterized these coefficients as evidence providing "strong support for the precision of WISC-V scores" (p. 39). Given these strong reliability coefficients, clinicians have been encouraged to interpret WISC-V^{CDN} score patterns, especially those at the factor index level (Wechsler, 2014b).

However, neither the method used to split subtests into equivalent scales nor the method used to compute reliability indices for composite scores were specifically identified by Wechsler (2014b). Critically, the assumptions underlying the split-half method were neither explicated nor satisfied. Internal consistency coefficients may be biased, either too low *or* too high, when those assumptions are violated, as they undoubtedly are in multidimensional tests like the WISC-V^{CDN} (Brown, 2015; Brunner et al., 2012; Raykov & Marcoulides, 2011). Given this reality, Geisinger (2013, p. 41) concluded that split-half reliability "should be seen primarily as a historical approach to estimating reliability . . . [and] there is simply no reason to use these procedures today." This opinion seems to be widely shared among measurement experts (Evers et al., 2013; Raykov & Marcoulides, 2011).

As an alternative to split-half and alpha internal consistency reliability estimates, model-based reliability estimates that make fewer and more realistic assumptions have been developed (Dunn, Baguley, & Brunsden, 2014; Reise, 2012). Critically, model-based estimates properly estimate reliability for multidimensional tests where item scales and factor loadings differ (Brunner et al., 2012; Hancock & Mueller, 2001). The omega (ω) family of coefficients (McDonald, 1999; Zinbarg, Revelle, Yovel, & Li, 2005) are the principal model-based reliability coefficients reported in current research. They replace the classical test theory hypothesis of true and error variance with the factor analytic conceptualization of common and unique variance. Especially for multidimensional tests like the WISC-V^{CDN}, omega "provides a better estimate for the composite score and thus should be used (Chen et al., 2012, p. 228).

There are several omega variants. The most general omega coefficient is omega total (ω), which is an "estimate of the proportion of variance in the unit-weighted total score attributable to all sources of common variance" (Rodriguez, Reise, & Haviland, 2016, p. 224). High ω values indicate a highly reliable multidimensional total composite score. ω can also be computed for each subscale score using the same logic. That is, the proportion of variance in each unit-weighted subscale score that can be attributed to a blend of general and group factor variance. Called omega subscale (ω_s), high values indicate a highly reliable multidimensional group composite score. Akin to coefficient alpha, ω and ω_s both reflect the systematic variance attributable to multiple common factors but neither can distinguish between the precision of the general factor versus the precision of the group factor (Brunner et al., 2012). There is no universally accepted guideline for coefficient alpha values sufficient for high-stakes decisions about individuals, but values in the .80–.90 range are commonly recommended (Thorndike & Thorndike-Christ, 2010; Wasserman & Bracken, 2013). Given their similarity, omega coefficients should meet the same standard as alpha coefficients.

The distinction between general and group factor variance can be made with omega hierarchical coefficients because they reflect variance attributable to a single factor independent of all other factors and are therefore measures of the precision with which a score assesses a single construct. When applied to the general factor, omega hierarchical (ω_h) is the ratio of the variance of the general factor compared to the total test variance and "reflects the percentage of systematic variance in unit-weighted total scores that can be attributed to the individual differences on the general factor" (Rodriguez et al., 2016, p. 224). Called omega hierarchical subscale (ω_{hs}) when applied to group factors, this index identifies the proportion of variance in the group factor score that is solely accounted for by its intended construct. If ω_{hs} is low relative to ω_s, most of the reliable variance of that group factor score is due to the general factor, which precludes meaningful interpretation of that group factor score as an unambiguous indicator of the target construct (Rodriguez et al., 2016). In contrast, a robust ω_{hs} coefficient suggests that most of the reliable variance of that group factor score is independent of the general factor, and clinical interpretation of an examinee's strengths and weaknesses beyond the general factor can be conducted (Brunner et al., 2012; DeMars, 2013; Reise, 2012). There is no empirically based guideline for acceptable levels of omega hierarchical coefficients for individual clinical decisions, but it has been suggested that they should, at a minimum, exceed .50, although .75 would be preferred (Reise, 2012).

Goals

Given the foregoing discussion, the evidence provided by Wechsler (2014b) regarding the reliability and validity of the WISC-V^{CDN} is open to question. However, competent psychological practice demands strong supportive evidence of reliability and validity before any test, including the WISC-V^{CDN}, can be used to make high-stakes decisions about vulnerable children (AERA, APA, & NCME, 2014). Consequently, the factor structure of the WISC-V^{CDN} was exhaustively analyzed to identify an appropriate scoring structure; and modern model-based estimates of reliability for the WISC-V^{CDN} were computed for a variety of models.

Method

Participants

Participants were the WISC-V^{CDN} standardization sample of children aged 6 years–16 years of age. Norming was conducted in 2013–2014 and included 880 children, stratified by age, sex, race and ethnicity, parent education level, and geographic region. The sample appeared to be representative of English-speaking Canadian children (see Wechsler, 2014b for full details).

Instruments

The WISC-V^{CDN} is a norm-referenced, individually administered intelligence battery appropriate for children aged 6 through 16 years. According to the WISC-V^{CDN} manual, CHC theory as well as neurodevelopmental research and clinical utility were considered in its development and these frameworks can be utilized in the interpretation of WISC-V^{CDN} scores.

The WISC-V^{CDN} contains 10 primary and 6 secondary subtests (M = 10, SD = 3). The 5 CHC factor index scores (M = 100, SD = 15) are computed from the 10 primary subtests: Similarities (SI) and Vocabulary (VO) create the Verbal Comprehension Index (VCI); Block Design (BD) and Visual Puzzles (VP) subtests create the Visual Spatial Index (VS); Matrix Reasoning (MR) and Figure Weights (FW) subtests create the Fluid Reasoning Index (FR); Digit Span (DS) and Picture Span (PS) subtests create the Working Memory Index (WMI); and Coding (CD) and Symbol Search (SS) subtests create the Processing Speed Index (PSI). The Full Scale IQ (FSIQ; M = 100, SD = 15) is computed using only 7 primary subtests: SI, VO, BD, MR, FW, DS, and CD. The 6 secondary subtests are proposed to load onto the same factors as the primary subtests: Information (IN) and Comprehension (CO) on the VC factor; Picture Concepts (PC) on the FR factor; Arithmetic (AR) and Letter-Number Sequencing (LN) on the WM factor; and Cancellation (CA) on the PS factor.

Wechsler (2014b) provides considerable information on the reliability and validity of WISC-V^{CDN} scores. For example, the average split-half reliability of the FSIQ for the total standardization sample was .96, whereas the average reliability of factor index scores ranged from .88 (PSI) to .93 (FRI). The reliability of subtest scores ranged from .81 (SS) to FW (.94). Concurrent validity was supported by a comparison of WISC-V^{CDN} scores to other cognitive and achievement tests. Convergent and discriminant validity was supported by studies of WISC-V^{CDN} scores among clinical groups. Factorial validity evidence was presented via a series of CFA with the final structural model adhering to a CHC framework (illustrated in Figure 1).

Analyses

Correlations, means, and standard deviations of the 16 WISC-V^{CDN} primary and secondary subtests for the total standardization sample were extracted from Table 4.1 of Wechsler (2014b). All CFA were conducted with Mplus 7.4 (Muthén & Muthén, 2015) from covariance matrices using the maximum likelihood estimator. Latent variable scales were identified by setting a reference indicator in higher-order models and by setting the variance of latent variables in bifactor models (Brown, 2015). Parameter estimates were constrained to equality in models with only two indicators per factor (Gignac, 2007).

Models

The evaluated models were duplicates of those specified by Wechsler (2014b, p. 50) and are identified in Table 1. Bifactor variants of simple structure models were also included to allow a comparison of alternative models not tested by Wechsler (2014b). Global model fit was evaluated with the chi-square likelihood ratio, comparative fit index (CFA), Tucker-Lewis index (TLI), standardized root mean square residual (SRMR), root mean square error of approximation (RMSEA), and Akaike's information criterion (AIC). Given the large sample size, it was expected that the chi-square likelihood test of exact fit would be rejected (Brown, 2015). Accordingly, global approximate fit measures that consider absolute (SRMR and RMSEA) and relative (CFI, TLI) fit as well as parsimony (RMSEA, AIC) were relied on to assess alternative models (Gignac, 2007; Loehlin & Beaujean, 2017). Based on prior research and expert suggestions (Hu & Bentler, 1999), good model fit required TLI/CFI ≥ .95 as well as SRMR and RMSEA ≤ .06. The AIC was used to compare the global fit of alternative models, with the lowest AIC value indicating the best model (Akaike, 1987). Meaningful differences between well-fitting models were also evaluated using ΔCFI/TLI > .01, ΔRMSEA > .015 (Chen, 2007; Cheung & Rensvold, 2002; Gignac, 2007), and ΔAIC > 10 (Burnham & Anderson, 2004). Given that global fit indices are averages that can mask areas of local misfit (McDonald & Ho, 2002) and potentially invalidate a model, parameter estimates were scrutinized to ensure that they made statistical and substantive sense (Brown, 2015).

Results

Global fit measures for all tested models are reported in Table 2. Models with fewer than four group factors failed to achieve good model fit, whereas models with four and five group factors generally achieved good global fit. For models with four group factors, the bifactor version of the traditional Wechsler model was superior (see Figure 3). For the models with five group factors in a CHC structure, a relaxed bifactor version of model 5a exhibited the best global fit (see Figure 4). The initial bifactor version of model 5a was improper, exhibiting a negative variance estimate for the FR

Table 1. Alternative structural models for the WISC-V^{CDN} with 16 primary and secondary subtests.

Model	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
1	SI-VO-IN-CO-AR-DS-LN- BD-VP-MR-FW-PC-PS-CD-SS-CA				
2	SI-VO-IN-CO-AR-DS-LN	BD-VP-MR-FW-PC-PS-CD-SS-CA	–	–	–
3	SI-VO-IN-CO-AR-DS-LN	BD-VP-MR-FW-PC-PS	CD-SS-CA	–	–
4a	SI-VO-IN-CO	BD-VP-MR-FW-PC	AR-DS-PS-LN	CD-SS-CA	–
4b	SI-VO-IN-CO	MR-FW-PC-AR-DS-PS-LN	BD-VP	CD-SS-CA	–
4c	SI-VO-IN-CO	$\underline{\text{AR}}$-BD-VP-MR-FW-PC	$\underline{\text{AR}}$-DS-PS-LN	CD-SS-CA	–
4d	$\underline{\text{AR}}$-SI-VO-IN-CO	$\underline{\text{AR}}$-BD-VP-MR-FW-PC	$\underline{\text{AR}}$-DS-PS-LN	CD-SS-CA	–
5a	SI-VO-IN-CO	BD-VP	MR-FW-PC	AR-DS-PS-LN	CD-SS-CA
5b	SI-VO-IN-CO	BD-VP	AR-MR-FW-PC	DS-PS-LN	CD-SS-CA
5c	SI-VO-IN-CO	BD-VP	$\underline{\text{AR}}$-MR-FW-PC	$\underline{\text{AR}}$-DS-PS-LN	CD-SS-CA
5d	$\underline{\text{AR}}$-SI-VO-IN-CO	BD-VP	MR-FW-PC	$\underline{\text{AR}}$-DS-PS-LN	CD-SS-CA
5e	$\underline{\text{AR}}$-SI-VO-IN-CO	BD-VP	$\underline{\text{AR}}$-MR-FW-PC	$\underline{\text{AR}}$-DS-PS-LN	CD-SS-CA

Note. SI = Similarities, VO = Vocabulary, IN = Information, CO = Comprehension, DS = Digit Span, LN = Letter-Number Sequencing, PS = Picture Span, AR = Arithmetic, MR = Matrix Reasoning, FW = Figure Weights, PC = Picture Concepts, BD = Block Design, VP = Visual Puzzles, CD = Coding, SS = Symbol Search, and CA = Cancellation. Complex loadings in each model underlined.

Table 2. Fit statistics for WISC-V^{CDN} 16 primary and secondary subtests for the total standardization sample (N = 880).

Model[a]	χ^2	*df*	CFI	TLI	SRMR	RMSEA	RMSEA 90% CI	AIC
1 (*g*)	1026.8	104	.829	.803	.062	.100	.095–.106	65914
2 (V-P)[b]	902.2	103	.852	.828	.060	.094	.088–.100	65791
3 (V-P-PS)	635.3	101	.901	.883	.048	.078	.072–.083	65528
4a (VC-VS-WM-PS)	364.9	100	.951	.941	.039	.055	.049–.061	65260
4a Bifactor	214.8	88	.977	.968	.026	**.040**	.034–.047	65133
4b (VC-VS-WM-PS)	390.6	100	.946	.936	.039	.057	.052–.064	65285
4c (VC-VS-WM-PS)	325.8	99	.958	.949	.037	.051	.045–.057	65223
4d (VC-VS-WM-PS)	305.5	98	.962	.953	.036	.049	.043–.055	65204
5a (VC-VS-FR-WM-PS)	330.8	99	.957	.948	.038	.052	.046–.058	65228
5a Bifactor[b]	Covariance matrix not positive definite. Negative variance estimate for Matrix Reasoning subtest.							
5a Bifactor Modified	208.1	87	**.978**	**.969**	**.025**	**.040**	.033–.047	**65129**
5b (VC-VS-FR-WM-PS)	Covariance matrix not positive definite. Negative variance estimate for FR factor.							
5b Bifactor[b]	231.4	89	.974	.965	.028	.043	.036–.049	65148
5c (VC-VS-FR-WM-PS)	Covariance matrix not positive definite. Negative variance estimate for FR factor.							
5d (VC-VS-FR-WM-PS)	277.9	98	.967	.959	.035	.046	.039–.052	65177
5e (VC-VS-FR-WM-PS)	267.5	97	.968	.961	.035	.045	.038–.051	65168

Note. CFI = Comparative Fit Index, TLI = Tucker–Lewis Index, SRMR = Standardized Root Mean Square, RMSEA = Root Mean Square Error of Approximation, AIC = Akaike's Information Criterion, *g* = general intelligence, V = Verbal, P = Performance, PS = Processing Speed, VC = Verbal Comprehension, VS = Visual Spatial, WM = Working Memory, FR = Fluid Reasoning. [a]Model numbers and letters correspond to those reported in the WISC–V^{CDN} *Manual* (plus bifactor variants of simple structure models that were added for this study) and are higher-order models (unless otherwise specified) when more than one first-order factor was specified. [b]Models with only two indicators were constrained to equality for identification. Best fit indicators in bold. Indices not meaningfully different (ΔCFI and ΔTLI < .01, ΔRMSEA > .015, ΔAIC ≤ 10) from best fit shaded.

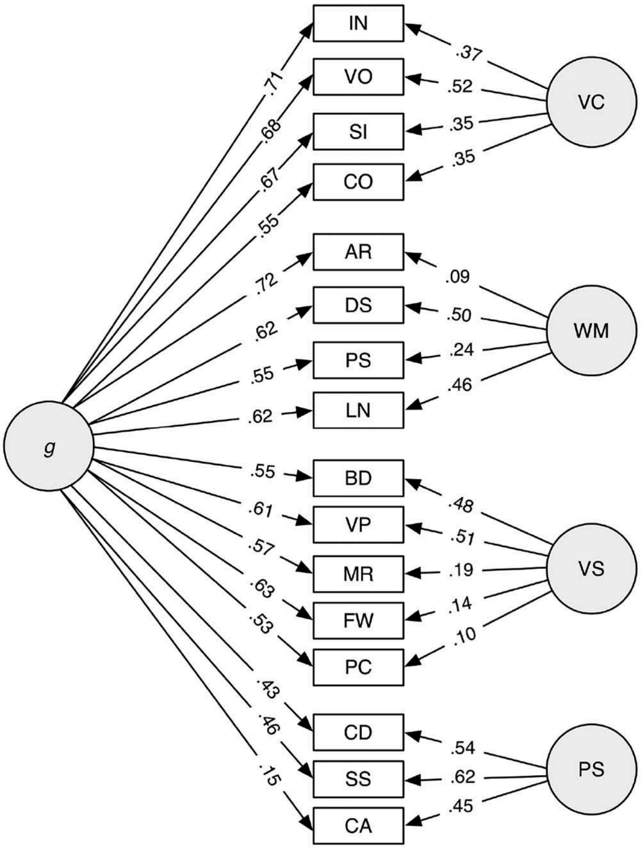

Figure 3. Standardized bifactor structure with best fit among the WISC-V^{CDN} primary and secondary subtests.

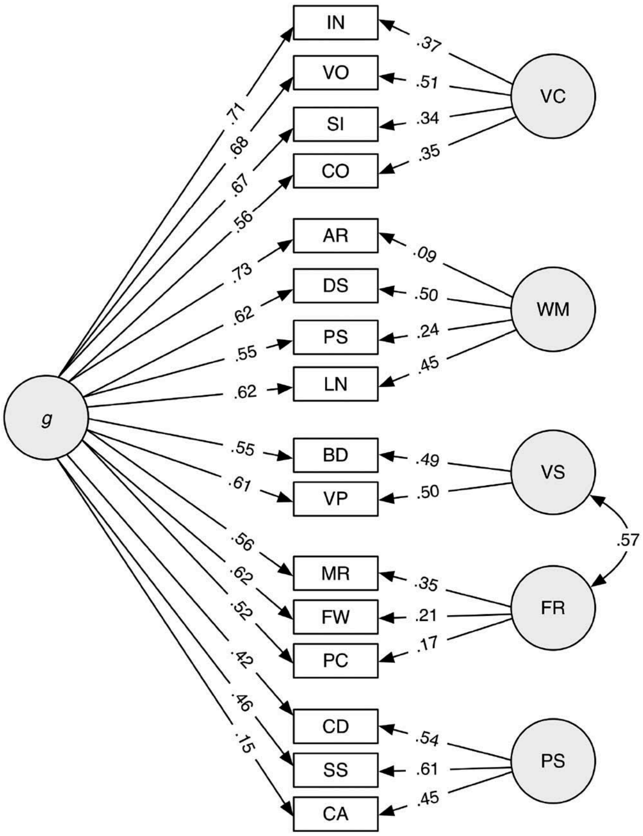

Figure 4. Modified Model 5a bifactor structure of the WISC-V^{CDN} primary and secondary subtests.

factor. When the FR and VS factors were allowed to correlate, the relaxed bifactor CHC model converged appropriately.

Given that evaluating models exclusively on the basis of global fit is insufficient, all models with good fit were also scrutinized for size of parameters, statistical significance of parameters, and interpretability (Bowen & Guo, 2012; Brown, 2015; Kline, 2016). Three models manifested negative error variances (see Table 2), which are statistically improper solutions that signal model misspecification

(McDonald, 2004). Although negative parameter values can be fixed to a small positive value to achieve a solution, “these remedial strategies are not recommended” (Brown, 2015, p. 167) because the resulting parameter estimates are likely to be biased and “should not be trusted” (Kline, 2016, p. 237). Consequently, those three offending models were removed from further consideration.

All higher-order models with CHC structure and complex loadings were marked by either improper solutions (5b and 5c) or with FR loadings on the general intelligence factor at such high levels (e.g., .99 for models 5d and 5e with attendant nonsignificant variance estimates) as to indicate that those factors were empirically redundant (Bowen & Guo, 2012; Le et al., 2010). Among the bifactor models with five group factors, only a modification of the bifactor version of model 5a generated a proper solution (see Figure 4). Thus, most models with five group factors exhibited good global fit but were invalidated by parameters that were statistically or substantively improper.

Based on global fit and simple structure, the bifactor version of model 4a was the superior model. All parameters were statistically significant, none were out-of-range, and all were substantively meaningful. However, it had two weaknesses: (a) several subtests had weak loadings on their group factor (AR at .09 onto WM and three indicators of VS ranging from .10 to .19) and (b) there was a substantial difference in loading strength between BD and VP subtests (*Md* = .50) when compared to MR, FW, and PC subtests (*Md* = .14). Thus, merging the FR and VS factors revealed strain. Of note, the loading of AR on the WM factor was low but statistically significant, whereas its loading was nonsignificant when allowed to load onto VC or PR instead of WM.

Given the parsimony afforded by its simple structure, the bifactor Wechsler model illustrated in Figure 3 was used for variance decomposition and computation of model-based reliability coefficients. Sources of variance from that model are presented in Table 3. The general factor accounted for 33.8% of the total variance and 67.6% of the common variance. None of the group factors accounted for substantial portions of variance. In fact, the general factor accounted for more than twice the total and common variance of all four group factors combined. Two subtests, PC and CA, exhibited low communalities, indicating that they did not substantially contribute to any of the factors. Only two subtests (IN and AR) were good measures of *g*, whereas three subtests (CD, SS, and CA) were poor measures of *g* (Kaufman, 1994). CA was an especially poor measure of *g*, loading at only .15. The other 11 subtests were fair measures of *g* (i.e., loadings from 50 to .69). All together, the general and group factors accounted for 50% of the total variance leaving another 50% due to specific variance and error.

The omega coefficients for the bifactor Wechsler model are reported in Table 4. They indicate that the general, VC, VS, and WM unit-weighted factor scores were reasonably reliable (i.e., near the .80–.90 range) in the sense that they reflected the systematic variance attributable to multiple common factors. However, when the systematic variance attributable to a single target factor of interest was indexed via omega hierarchical coefficients, only the general factor exhibited good reliability (ω_h = .83), whereas the group factor coefficients were low (ω_{hs} = .15–.48), suggesting that “much of the reliable variance of the subscale scores can be attributable to the general factor, and not what is unique to the group factors” (Rodriguez et al., 2016, p. 225). For example, 85% of the variance of the unit-weighted VCI score was attributable to both general and VC factors, whereas only 23% of the variance of that same unit-weighted VCI score was uniquely attributable to the VC factor. Another perspective on WISC-V^{CDN} reliability can be obtained by computing omega coefficients for a CHC higher-order model. Accordingly, the oblique results from model 5a were transformed into an orthogonal solution with the Schmid and Leiman (1957) method. Resulting omega coefficients are presented in Table 4. Omega coefficients for the general, VC, and WM unit-weighted factor scores were in the .80–.90 range but omega hierarchical coefficients were all below .50.

These reliability estimates are hypothetical because neither of these models represents the actual scoring structure of the WISC-V^{CDN}, which creates 5 CHC factors from 10 primary subtests. That oblique structure was also transformed into an orthogonal solution with the Schmid and Leiman (1957) method, with resultant omega coefficients presented in Table 4. Only the omega coefficient for the total score (.89) and omega hierarchical coefficient for the total score (.81) were sufficient for individual decisions. The omega coefficients for the total score in this model are also hypothetical because the WISC-V^{CDN} FSIQ is actually computed from only 7 subtests, not the 10 subtests needed to obtain factor index scores. When the 3 extraneous subtests were omitted from the primary subtest model, estimates of ω and ω_h were .85 and .77, respectively. Thus, the shortened FSIQ was reduced in reliability but still sufficiently precise for high-stakes individual decisions.

Discussion

Standardization sample data from the WISC-V^{CDN} were analyzed to investigate reliability and structural validity of its scores. Although Wechsler (2014b) preferred a complex higher-order CHC model, the new FR factor in that model was problematic because it

Table 3. Sources of variance for WISC-V^{CDN} standardization sample (N = 880) according to a bifactor Wechsler model.

	General		Verbal Comprehension		Visual Spatial		Working Memory		Processing Speed			
Subtest	b	b^2	b	b^2	b	b^2	b	b^2	b	b^2	h^2	u^2
Similarities	*.669*	.448	.346	.120							.567	.433
Vocabulary	*.676*	.457	.517	.267							.724	.276
Information	**.707**	.500	.372	.138							.638	.362
Comprehension	*.554*	.307	.350	.122							.429	.571
Block Design	*.550*	.303			.476	.227					.529	.471
Visual Puzzles	*.611*	.373			.506	.256					.629	.371
Matrix Reasoning	*.569*	.324			.189	.036					.359	.641
Figure Weights	*.629*	.396			.135	.018					.414	.586
Picture Concepts	*.529*	.280			.103	.011					.290	.710
Arithmetic	**.724**	.524					.093	.009			.533	.467
Digit Span	*.621*	.386					.497	.247			.633	367
Picture Span	*.552*	.305					.238	.057			.361	.639
Letter–Number Sequencing	*.621*	.386					.455	.207			.593	.407
Coding	.425	.181							.541	.293	.473	.527
Symbol Search	.464	.215							.616	.379	.595	.405
Cancellation	.150	.022							.454	.206	.229	.771
Total Variance		.338		.040		.034		.032		.055	.500	.500
Common Variance		.676		.081		.068		.065		.110		

Note. b = standardized loading of subtest on factor; b^2 = variance explained in the subtest; h^2 = communality; u^2 = uniqueness. g loadings ≥ .70 are considered good (bold), from .50 to .69 are fair (italic), and < .50 are poor (Kaufman, 1994).

Table 4. Omega reliability coefficients for standardization sample (N = 880) from alternative models.

	Bifactor Wechsler 16 Subtests		CHC 16 Subtests		10 Primary Subtests	
Factor Score	ω/ω_s	ω_h/ω_{hs}	ω/ω_s	ω_h/ω_{hs}	ω/ω_s	ω_h/ω_{hs}
General	**.919**	**.830**	**.918**	**.839**	**.888**	**.805**
Verbal Comprehension	**.850**	.230	**.848**	.251	.773	.252
Working Memory	**.809**	.167	**.794**	.178	.632	.180
Visual Spatial	.788	.152	.737	.255	.736	.225
Fluid Reasoning	–	–	.643	.036	.593	.042
Processing Speed	.683	.483	.676	.452	.696	.449

Note. ω and ω_s = omega of general and group factors, respectively; ω_h and ω_{hs} = omega hierarchical of general and group factors, respectively. Omega coefficients should exceed ~.80 for decisions about individuals (Thorndike & Thorndike-Christ, 2010; Wasserman & Bracken, 2013). At a minimum, omega hierarchical coefficients should exceed .50, although .75 would be preferred (Reise, 2012). Coefficients meeting minimum standards are in bold.

produced negative variance estimates, empirically redundant factors, or low unique reliability estimates. At best, it lacked discriminant validity (Le et al., 2010). An alternative bifactor model with four group factors and one general factor akin to the traditional Wechsler structure was judged to be a good representation of the structure of the WISC-V^{CDN}. Alternatively, a bifactor CHC model with correlated FR and VS factors exhibited good fit but poor discriminant validity and concomitant interpretational confounding (Stromeyer et al., 2015). These results are not surprising, given that previous Wechsler scales as well as the U.S. version of the WISC-V have been found consistent with similar bifactor models (Canivez et al., 2016; Gignac & Watkins, 2013; Gomez, Vance, & Watson, 2017; Gustafsson & Undheim, 1996; Reynolds & Keith, 2017; Styck & Watkins, 2016; Watkins, 2006; Watkins, Canivez, James, James, & Good, 2013).

The merits of bifactor versus higher-order models have recently received considerable attention. Murray and Johnson (2013) found that fit indices are biased in favor of the bifactor model when there are unmodeled complexities (e.g., minor loadings of indicators on multiple factors). Morgan, Hodge, Wells, and Watkins (2015) analyzed simulations of bifactor and higher-order models and confirmed that both models exhibited good model fit regardless of true structure. More recently, Mansolf and Reise (2017) confirmed that bifactor and higher-order models could not be distinguished by fit indices and admitted that there is, at present, no technical solution to this dilemma.

Given that bifactor and higher-order models cannot be confidently differentiated, it seems reasonable to require "a parsimonious, substantively meaningful model that fits observed data adequately well" (MacCallum & Austin, 2000, p. 218) that fulfills the purpose of measurement. Murray and Johnson (2013) suggested that either model would provide a good estimate of general intelligence but "if 'pure' measures of specific abilities are required then bi-factor model factor scores should be preferred to those from a higher-order model" (p. 420). This logic has been endorsed by other measurement specialists (Brunner et al., 2012; DeMars, 2013; Morin, Arens, Tran, & Caci, 2016; Reise, 2012; Reise, Bonifay, & Haviland, 2013; Rodriguez et al., 2016). Given that scores from the WISC-V^{CDN} will likely be used by psychologists to provide an estimate of general ability *and* to identify interventions based on cognitive strengths and weaknesses as operationalized through the factor index scores (Wechsler, 2014b), bifactor model factor scores would be preferred (Murray & Johnson, 2013).

As predicted by Murray and Johnson (2013), all models considered in this study produced reasonable estimates of general ability. However, omega coefficients computed from both bifactor and higher-order models demonstrated that reliable variance of a WISC-V^{CDN} factor index score was primarily due to the general factor, *not* the group factor (see Table 4). An additional consideration for clinical interpretation is that approximately 50% of the total variance of WISC-V^{CDN} scores was due to error and specific variance. Therefore, to interpret factor index scores "as representing the precise measurement of some latent variable that is unique or different from the general factor, clearly, is misguided" (Rodriguez et al., 2016, p. 225).

"The ultimate responsibility for appropriate test use and interpretation lies predominantly with the test user" (AERA, APA, & NCME, 2014, p. 141). This study demonstrated that psychologists can be reasonably confident in using the FSIQ score for clinical decisions but should be extremely cautious in using the factor index scores to make decisions about individuals. Factor index scores represent a blend of general *and* group abilities as well as error and usually provide little information beyond that provided by the general factor (Beaujean, Parkin, & Parker, 2014; Canivez, 2016; Cucina & Howardson, 2017). Interpretation of factor index scores should also be informed by external validity evidence (AERA, APA, & NCME, 2014; Hummel, 1998; Wasserman & Bracken, 2013). DeMars (2013) predicted that differential validity would be impaired by scores with low precision. That prediction has been supported in many studies of external validity (Carroll, 2000). For example, there is little evidence to support the proposition that factor score differences validly inform diagnosis or treatment (Braden & Shaw, 2009; Burns, 2016; Kearns & Fuchs, 2013; Kranzler, Benson, et al., 2016; Kranzler, Floyd, Benson, Zaboski, & Thibodaux, 2016; Reschly, 1997; Restori, Gresham, & Cook, 2008). Likewise, multiple studies have found little incremental validity for Wechsler factor index scores beyond the FSIQ when predicting academic achievement (Benson et al., 2016; Canivez, 2013; Canivez, Watkins, James, Good, & James, 2014; Glutting, Watkins, Konold, & McDermott, 2006). Also, the predictive power of FSIQ scores is not diminished by variability among factor scores (Daniel, 2007; McGill, 2016; Watkins, Glutting, & Lei, 2007). The cumulative weight of this reliability and validity evidence suggests that psychologists should focus their interpretive efforts at the general factor level, and exercise extreme caution when using group factor scores to make decision about individuals.

About the authors

Marley W. Watkins, PhD, is a Non-Resident Scholar in the Department of Educational Psychology at Baylor University. His research interests include professional issues, the psychometrics of assessment and diagnosis, and individual differences.

Stefan C. Dombrowski, is Professor and Director of the School Psychology Program at Rider University. His research has focused on such topics as structural validity, prenatal exposures, children's mental health, and psychoeducational assessment.

Gary L. Canivez, PhD, is Professor of Psychology at Eastern Illinois University. His research interests include psychometric reliability and validity investigations of measures of intelligence, achievement, psychopathology, and test bias.

ORCID

Marley W. Watkins http://orcid.org/0000-0001-6352-7174
Gary L. Canivez http://orcid.org/0000-0002-5347-6534

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*, 317–332.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Australian Psychological Society. (2009). *Guidelines for psychological assessment and the use of psychological tests*. Melbourne, Australia: Author.

Beaujean, A. A. (2015a). Adopting a new test edition: Psychometric and practical considerations. *Research and Practice in the Schools*, *3*, 51–57.

Beaujean, A. A. (2015b). John Carroll's views on intelligence: Bi-factor vs. higher-order models. *Journal of Intelligence*, *3*, 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

Beaujean, A. A., Parkin, J., & Parker, S. (2014). Comparing Cattell-Horn-Carroll factor models: Differences between bifactor and higher order factor models in predicting language achievement. *Psychological Assessment*, *26*, 789–805. doi:10.1037/a0036745

Benson, N., Beaujean, A. A., & Taub, G. E. (2015). Using score equating and measurement invariance to examine the Flynn effect in the Wechsler adult intelligence scale. *Multivariate Behavioral Research*, *50*, 398–415. doi:10.1080/00273171.2015.1022642

Benson, N. F., Kranzler, J. H., & Floyd, R. G. (2016). Examining the integrity of measurement of cognitive abilities in the prediction of achievement: Comparisons and contrasts across variables from higher-order and bifactor models. *Journal of School Psychology*, *58*, 1–19. doi:10.1016/j.jsp.2016.06.001

Bowen, N. K., & Guo, S. (2012). *Structural equation modeling*. New York, NY: Oxford University Press.

Braden, J. P., & Shaw, S. R. (2009). Intervention validity of cognitive assessment: Knowns, unknowables, and unknowns. *Assessment for Effective Intervention*, *34*, 106–115.

British Psychological Society. (2007). *Psychological testing: A user's guide*. Leicester, UK: Author.

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

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

Burns, M. K. (2016). Effect of cognitive processing assessments and interventions on academic outcomes. *Can 200 Studies Be Wrong? Communiqué*, *44*(5), 1, 26–29.

Canadian Psychological Association. (2000). *Canadian code of ethics for psychologists* (3rd ed.). Ottawa, Canada: Author.

Canadian Psychological Association. (2007). *Professional practice guidelines for school psychologists in Canada*. Ottawa, Canada: Author.

Canivez, G. L. (2013). Incremental criterion validity of WAIS-IV factor index scores: Relationships with WIAT-II and WIAT-III subtest and composite scores. *Psychological Assessment*, *25*, 484–495. doi:10.1037/a0032092

Canivez, G. L. (2014). Construct validity of the WISC-IV with a referred sample: Direct versus indirect hierarchical structures. *School Psychology Quarterly*, *29*, 38–51. doi:10.1037/spq0000032

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). Gottinger, Germany: Hogrefe.

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

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

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*. doi:10.1037/pas0000358

Canivez, G. L., Watkins, M. W., James, T., Good, R., & James, K. (2014). Incremental validity of WISC-IV^{UK} factor index scores with a referred Irish sample: Predicting performance on the WIAT-II^{UK}. *British Journal of Educational Psychology*, *84*, 667–684. doi:10.1111/bjep.12056

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

Carroll, J. B. (2000). Commentary on profile analysis. *School Psychology Quarterly*, *15*, 449–456.

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

Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J.-P., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. *Journal of Personality*, *80*, 219–251. doi:10.1111/j.1467-6494.2011.00739.x

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

Cormier, D. C., Kennedy, K. E., & Aquilina, A. M. (2016). Test review: Wechsler, D. (2014). Wechsler Intelligence Scale for Children–Fifth Edition: Canadian (WISC-V^{CDN}). *Canadian Journal of School Psychology*, *31*, 322–334.

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*. doi:10.1037/pas0000389

Daniel, M. H. (2007). "Scatter" and the construct validity of FSIQ: Comment on Fiorello et al. (2007). *Applied Neuropsychology*, *14*, 291–295.

DeMars, C. E. (2013). A tutorial on interpreting bifactor model scores. *International Journal of Testing*, *13*, 354–378. doi:10.1080/15305058.2013.799067

DiStefano, C., & Hess, B. (2005). Using confirmatory factor analysis for construct validation: An empirical review. *Journal of Psychoeducational Assessment*, *23*, 225–241.

Dombrowski, S. C., Canivez, G. L., Watkins, M. W., & Beaujean, A. 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. doi:10.1016/j.intell.2015.10.009

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*, 399–412. doi:10.1111/bjop.12046

Evers, A., Hagemeister, C., Høstmaelingen, P., Lindley, P., Muñiz, J., & Sjöberg, A. (2013). *EFPA review model for the description and evaluation of psychological and educational tests*. Brussels, Belgium: European Federation of Psychologists' Associations.

Evers, A., Muñiz, J., Bartram, D., Boben, D., Egeland, J., Fernández-Hermida, J. R., & Urbánek, T. (2012). Testing practices in the 21st century: Developments and European psychologists' opinions. *European Psychologist*, *17*, 300–319. doi:10.1027/1016-9040/a000102

Frazier, T. W., & Youngstrom, E. A. (2007). Historical increase in the number of factors measured by commercial tests of cognitive ability: Are we overfactoring? *Intelligence*, *35*, 169–182. doi:10.1016/j.intell.2006.07.002

Geisinger, K. F. (2013). Reliability. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology: Test theory and testing and assessment in industrial and organizational psychology* (Vol. 1, pp. 21–42). Washington, DC: American Psychological Association.

Gignac, G. E. (2007). Multi-factor modeling in individual differences research: Some recommendations and suggestions. *Personality and Individual Differences*, *42*, 37–48. doi:10.1016/j.paid.2006.06.019

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

Glutting, J. J., Watkins, M. W., Konold, T. R., & McDermott, P. A. (2006). Distinctions without a difference: The utility of observed versus latent factors from the WISC-IV in estimating reading and math achievement on the WIAT-II. *Journal of Special Education*, *40*, 103–114.

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*, 496–508. doi:10.1037/a0030676

Gomez, R., Vance, A., & Watson, S. (2017). Bifactor model of WISC-IV: Applicability and measurement invariance in low and normal IQ groups. *Psychological Assessment*. doi:10.1037/pas0000369

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

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.

Hoyle, R. H. (2000). Confirmatory factor analysis. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 465–497). New York, NY: Academic Press.

Hsu, C. T., Huang, C. H., & Cheng, C. C. (2009). Practice survey of clinical psychologists in Taiwan. *Research in Applied Psychology*, *41*, 43–55.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55. doi:10.1080/10705519909540118

Hummel, T. J. (1998). The usefulness of tests in clinical decisions. In J. W. Lichtenberg & R. K. Goodyear (Eds.), *Scientist-practitioner perspectives on test interpretation* (pp. 59–112). Boston, MA: Allyn & Bacon.

International Test Commission. (2001). International guidelines for test use. *International Journal of Testing*, *1*, 93–114.

Kaufman, A. S. (1994). *Intelligent testing with the WISC–III*. New York, NY: Wiley.

Kearns, D. M., & Fuchs, D. (2013). Does cognitively focused instruction improve the academic performance of low-achieving students? *Exceptional Children*, *79*, 263–290.

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

Kranzler, J. H. (2016). Current practices and future directions for the assessment of child and adolescent intelligence in schools around the world. *International Journal or School & Educational Psychology*, *4*, 213–214. doi:10.1080/21683603.2016.1166762

Kranzler, J. H., Benson, N., & Floyd, R. G. (2016). Intellectual assessment of children and youth in the United States of American: Past, present, and future. *International Journal of School & Educational Psychology*, *4*, 276–282. doi:10.1080/21683603.2016.1166759

Kranzler, J. H., Floyd, R. G., Benson, N., Zaboski, B., & Thibodaux, L. (2016). Cross-battery assessment pattern of strengths and weaknesses approach to identification of specific learning disorders: Evidence-based practice or pseudoscience? *International Journal of School & Educational Psychology*, *4*, 146–157. doi:10.1080/21683603.2016.1192855

Krishnamurthy, R., VandeCreek, L., Kaslow, N. J., Tazeau, Y. N., Miville, M. L., Kerns, R., & Benton, S. A. (2004). Achieving competency in psychological assessment: Directions for education and training. *Journal of Clinical Psychology*, *60*, 725–739. doi:10.1002/jclp.20010

Le, H., Schmidt, F. L., Harter, J. K., & Lauver, K. J. (2010). The problem of empirical redundancy of constructs in organizational research: An empirical investigation. *Organizational Behavior and Human Decision Processes*, *112*, 112–125. doi:10.1016/j.obhdp.2010.02.003

Lei, P.-W., & Wu, Q. (2012). Estimation in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 164–180). New York, NY: Guilford.

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling*, *13*, 59–72.

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

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology*, *51*, 201–226.

Mansolf, M., & Reise, S. P. (2017). When and why the second-order and bifactor models are distinguishable. *Intelligence*, *61*, 120–129. doi:10.1016/j.intell.2017.01.012

McDonald, R. P. (1999). *Test theory: A unified approach*. Mahwah, NJ: Erlbaum.

McDonald, R. P. (2004). Respecifying improper structures. *Structural Equation Modeling*, *11*, 194–209.

McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*, 64–82. doi:10.1037//1082-989X.7.1.64

McGill, R. J. (2016). Invalidating the full scale IQ score in the presence of significant factor score variability: Clinical acumen or clinical illusion? *Archives of Assessment Psychology*, *6*, 33–63.

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

Millsap, R. E. (2007). Structural equation modeling made difficult. *Personality and Individual Differences*, *42*, 875–881. doi:10.1016/j.paid.2006.09.021

Morgan, G. B., Hodge, K. J., Wells, K. E., & Watkins, M. W. (2015). Are fit indices biased in favor of bi-factor models in cognitive ability research?: Acomparison of fit in correlated factors, higher-order, and bi-factor models via Monte Carlo simulations. *Journal of Intelligence*, *3*, 2–20. doi:10.3390/jintelligence3010002

Morin, A. J. S., Arens, A. K., Tran, A., & Caci, H. (2016). Exploring sources of construct-relevant multidimensionality in psychiatric measurement: A tutorial and illustration using the composite scale of morningness. *International Journal of Methods in Psychiatric Research*, *25*, 277–288. doi:10.1002/mpr.1485

Murray, A. L., & Johnson, W. (2013). The limitations of model fit in comparing bi-factor versus higher-order models of human cognitive ability structure. *Intelligence*, *41*, 407–422. doi:10.1016/j.intell.2013.06.004

Muthén, B. O., & Muthén, L. K. (2015). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*, 190–204.

Na, S. D., & Burns, T. G. (2016). Wechsler Intelligence Scale for Children–V: Test review. *Applied Neuropsychology: Child*, *5*, 156–160. doi:10.1080/21622965.2015.1015337

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175–220.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York, NY: Taylor & Francis.

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., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, *92*, 544–559. doi:10.1080/00223891.2010.496477

Reschly, D. J. (1997). Diagnostic and treatment utility of intelligence tests. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 437–456). New York, NY: Guilford.

Restori, A. F., Gresham, F. M., & Cook, C. R. (2008). "Old habits die hard": Past and current issues pertaining to response-to-intervention. *California School Psychologist*, *13*, 67–78.

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

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment*, *98*, 223–237. doi:10.1080/00223891.2015.1089249f

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

Schneider, W. J., & McGrew, K. S. (2012). The model of intelligence. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment* (3rd ed., pp. 99–144). New York, NY: Guilford Press.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, *46*, 561–584.

Strauss, E., Spreen, O., & Hunter, M. (2000). Implications of test revisions for research. *Psychological Assessment*, *12*, 237–244. doi:10.1037//1040-3590.12.3.237

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*, 491–520. doi:10.1177/0149206314551962

Styck, K. M., & Watkins, M. W. (2016). Structural validity of the WISC-IV for students with learning disabilities. *Journal of Learning Disabilities*, *49*, 216–224. doi:10.1177/0022219414539565

Thorndike, R. M., & Thorndike-Christ, T. (2010). *Measurement and evaluation in psychology and education* (8th ed.). New York, NY: Pearson.

Thurstone, L. L. (1947). *Multiple-factor analysis*. Chicago, IL: University of Chicago Press.

Wasserman, J. D., & Bracken, B. A. (2013). Fundamental psychometric considerations in assessment. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of psychology: Assessment psychology* (Vol. 10, 2nd ed., pp. 50–81). Hoboken, NJ: Wiley.

Watkins, M. W. (2006). Orthogonal higher order structure of the Wechsler Intelligence Scale for Children–Fourth Edition. *Psychological Assessment*, *18*, 123–125. doi:10.1037/1040-3590.18.1.123

Watkins, M. W., & Beaujean, A. A. (2014). Bifactor structure of the Wechsler Preschool and Primary Scale of Intelligence–Fourth Edition. *School Psychology Quarterly*, *29*, 52–63. doi:10.1037/spq0000038

Watkins, M. W., Canivez, G. L., James, T., James, K., & Good, R. (2013). Construct validity of the WISC-IVUK with a large referred Irish sample. *International Journal of School & Educational Psychology*, *1*, 102–111. doi:10.1080/21683603.2013.794439

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

Wechsler, D. (2014a). *Wechsler Intelligence Scale for Children–Fifth Edition: Canadian*. Toronto, Canada: Pearson Canada Assessment.

Wechsler, D. (2014b). *Wechsler Intelligence Scale for Children–Fifth Edition: Canadian manual*. Toronto, Canada: Pearson Canada Assessment.

Wechsler, D. (2014c). *Wechsler Intelligence Scale for Children–Fifth Edition*. Bloomington, MN: Pearson Clinical Assessment.

Widaman, K. F. (2012). Exploratory factor analysis and confirmatory factor analysis. In H. Cooper (Ed.), *APA handbook of research methods in psychology: Data analysis and research publication* (Vol. 3, pp. 361–389). Washington, DC: American Psychological Association.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω_h: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*, 123–133. doi:10.1007/s11336-003-0974-7