INTERNATIONAL JOURNAL OF SCHOOL & EDUCATIONAL PSYCHOLOGY
2019, VOL. 7, NO. 3, 150–164
https://doi.org/10.1080/21683603.2017.1414006

Routledge
Taylor & Francis Group

ARTICLE

Construct validity of the Spanish Version of the Wechsler Intelligence Scale for Children Fifth Edition (WISC-V^{Spain})

Javier Fenollar-Cortés[a] and Marley W. Watkins [b]

[a]Department of Psychology, Loyola University Andalucia, Sevilla, Spain; [b]Department of Educational Psychology, Baylor University, Waco, Texas, USA

ABSTRACT

The construct validity of the Spanish Version of the Wechsler Intelligence Scale for Children Fifth Edition (WISC-V^{Spain}) was investigated via confirmatory factor analysis. For all 15 subtests, the higher-order model preferred by Wechsler (2015b) contained five group factors but lacked discriminant validity. A bifactor model with five group factors and one general factor in a Cattell-Horn-Carroll framework exhibited good fit when the Fluid Reasoning and Visual Spatial group factors were allowed to correlate but was compromised by low discriminant validity with concomitant interpretation confounding. A bifactor model with four group factors and one general factor akin to the traditional Wechsler model also exhibited good global fit and afforded greater parsimony through simple structure and fewer factors. In both models, the general factor was predominant, accounting for around 35% of the total variance and 67% of the common variance and about twice the variance accounted for by all the group factors combined. Similar results were obtained when the 10 primary subtests were analyzed. For both 10- and 15-subtest analyses, results demonstrated that reliable variance of WISC-V^{Spain} factor index scores was primarily due to the general factor. 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; Spain; intelligence; validity; reliability

The administration and interpretation of standardized psychological tests to assess cognitive ability is a fundamental professional function of psychologists (Evers et al., 2017; Kranzler, 2016; Kranzler, Benson, & Floyd, 2016) and the Wechsler scales are the most popular standardized instruments used by those psychologists for that purpose (Oakland, Douglas, & Kane, 2016; Piotrowski, 2017; Wright et al., 2017). Therefore, versions of the Wechsler Intelligence Scale for Children, especially its most recent fifth edition, are widely applied across the globe (Wechsler, 2014a, 2014b, 2015a, 2016a, 2016b, 2016c).

When adapting a test from one country or culture to another, the International Test Commission (ITC, 2017) has suggested that evidence supporting the norms, reliability, and validity of the adapted version of the test be provided. Further, professional standards require that psychologists understand the psychometric principles of reliability and validity and accept responsibility for proper administration and accurate scoring of tests as well as for interpretation of test scores consistent with the evidence supporting their reliability and validity (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education [AERA, APA, & NCME], 2014; British Psychological Society, 2007; Evers et al., 2013; ITC, 2001). Consequently, competent psychological assessment of cognitive ability with a WISC is dependent upon evidence regarding the reliability and validity of its scores (Krishnamurthy et al., 2004).

WISC-V ^{Spain}

A revised and adapted version of the U.S. WISC-V (Wechsler, 2014a) was recently published in Spain: the *Escala de inteligencia de Wechsler para niños–V* (WISC-V^{Spain}; Wechsler, 2015a). Its scores cannot be assumed to be identical to those of its predecessor because the WISC-V^{Spain} was a major revision involving the addition of new subtests and factor index scores, deletion of subtests, and changes to the contents and instructions of all subtests (Wechsler,

CONTACT Javier Fenollar-Cortés jfenollar@uloyola.es Department of Psychology, Loyola University Andalucia, 41014, Sevilla, Spain.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/usep.
This article has been republished with minor changes. These changes do not impact the academic content of the article.

2015b). Nor can the WISC-V^{Spain} be assumed to be identical to the U.S. version because its adaptation involved the deletion of U.S. items as well as the addition of new items to several subtests and the omission of an entire U.S. subtest (i.e., Picture Concepts). Thus, its psychometric merits must be independently evaluated by prospective users (AERA, APA, & NCME, 2014; Beaujean, 2015a; ITC, 2016).

Psychometric information about the WISC-V^{Spain} was provided in its manual (Wechsler, 2015b). Additional information about the WISC-V^{Spain} might be obtained from independent test reviews but none have been published as yet. Thus, the only direct source of information about the reliability and validity of WISC-V^{Spain} scores comes from its publisher. Dependence on the opinion of test 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).

However, versions of the WISC-V have also been published in other countries, so analyses of those national scales might provide information about the validity of WISC-V^{Spain} scores, albeit indirectly. The publisher proposed a higher-order structure with a second-order general intelligence (*g*) factor being loaded by five first-order group factors for every national scale (e.g., Figure 1) but independent factor analyses of the U.S., UK, French, and Canadian normative scores have preferred a four-factor structure reminiscent of that found with the prior fourth edition of the WISC where the subtests purported to load onto a factor new to the WISC-V (fluid reasoning) combined with subtests previously found to measure visual-perceptual reasoning (Canivez, Watkins, & Dombrowski, 2016, 2017; Canivez, Watkins, & McGill, 2017; Dombrowski, Canivez, & Watkins, 2017; Lecerf & Canivez, 2017;Watkins, Dombrowski, & Canivez, 2017). Although an analysis of the U.S. normative data by Reynolds and Keith (2017) partially supported the publisher-preferred five-factor model, the fluid reasoning and visual-spatial reasoning factors were allowed to correlate to recognize "the nonverbal related nature of these two factors" (p. 38). However, there was no explanation provided to justify why the processing speed factor (also comprised of nonverbal content) was not allowed to correlate with the fluid and visual-spatial reasoning factors nor how interfactor correlations might compromise discriminant validity (Stromeyer, Miller, Sriramachandramurthy, & DeMartino, 2015). Factorial invariance of the publisher preferred five-factor structure has been reported (H. Chen, Zhang, Raiford, Zhu, & Weiss, 2015; Scheiber, 2016), but rival four-factor models were not investigated in those studies. Thus, independent researchers tend to disagree with the publisher regarding the constructs being measured by the WISC-V.

The ITC (2016) noted that determination of the factor structure of an adapted test is especially important. For the WISC-V^{Spain}, Wechsler (2015b) proposed a higher-order structure with a second-order general intelligence (*g*) factor being loaded by five first-order group factors which, in turn, were loaded by 15 primary and secondary subtests. This structure was established via confirmatory factor analysis (CFA) by Wechsler (2015b) and is illustrated in Figure 1. However, "CFA studies based upon weak theoretical

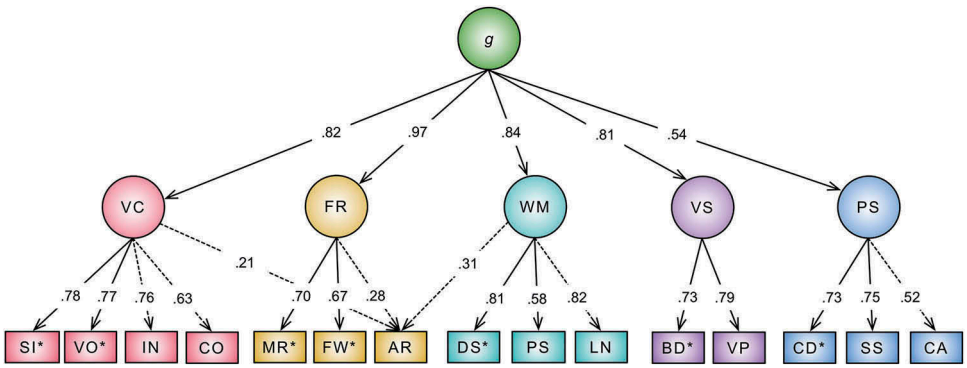

Figure 1. Standardized structure of the WISC-V^{Spain} proposed by Wechsler (2015b).

Note. SI = Similarities, VO = Vocabulary, IN = Information, CO = Comprehension, MR = Matrix Reasoning, FW = Figure Weights, AR = Arithmetic, DS = Digit Span, PS = Picture Span, LN = Letter-Number Sequencing, BD = Block Design, VP = Visual Puzzles, CD = Coding, SS = Symbol Search, CA = Cancellation, VC = Verbal Comprehension factor, FR = Fluid Reasoning factor, WM = Working Memory factor, VS = Visual Spatial factor, PS = Processing Speed factor, and *g* = General Intelligence. Solid lines are paths to primary subtests used to compute factor index scores and dotted lines are paths to secondary subtests. Asterisk (*) indicates subtests that contribute to the Full Scale IQ score.

perspectives, lack of testing alternative theoretical views, or insufficient evidence may not offer adequate support of construct validity" (DiStefano & Hess, 2005, p. 225). Additionally, CFA can demonstrate that a model is *consistent* with the data but "it does not *confirm* the veracity of the researcher's model" (Kline, 2016, p. 21).

Based on best-practice guidelines (Bowen & Guo, 2012; Brown, 2015; DiStefano & Hess, 2005; Kline, 2016; MacCallum & Austin, 2000; McDonald & Ho, 2002; Widaman, 2012), we have seven major concerns about the CFA methods reported by Wechsler (2015b). First, only higher-order models were tested. In the higher-order model, general intelligence (*g*) is seen as a superordinate factor having a direct effect on several group factors but an indirect effect on the measured variables (see top panel of Figure 2). In contrast, bifactor models (see bottom panel of Figure 2) conceptualize *g* as a breadth factor having direct effects on the measured variables (Reise, 2012). Carroll's (1993) cognitive model was incorporated into the Cattell-Horn-Carroll theory (CHC; Schneider & McGrew, 2012) that influenced development of the WISC-V^{Spain} (Wechsler, 2015b) and is most accurately represented by a bifactor model (Beaujean, 2015b). Thus, alternative conceptualizations of the factorial structure of the WISC-V^{Spain} must be tested to provide convincing support for a model (Brown, 2015).

Second, the method used to scale the latent variables in CFA models was not reported by Wechsler (2015b). All standard methods should produce identical degrees of freedom and model fit indices (Brown, 2015). Beaujean (2016) reproduced the analyses reported for the U.S. WISC-V (Wechsler, 2014a) and conjectured that an improperly modified effects-coding method was applied that understated the degrees of freedom. It appears that this modified effects-coding method was also applied to CFA analyses of the WISC-V^{Spain} standardization sample. This could impact fit indices and interpretation. Thus, it is unclear whether the models reported by Wechsler (2015b) were actually tested (Cortina, Green, Keeler, & Vandenberg, 2017). The consequences of using a nonstandard scaling method in the WISC-V^{Spain} analyses are unknown but should be explored through replication using standard scaling methods.

Third, Wechsler (2015b) used weighted least squares (WLS) to estimate CFA parameters whereas maximum

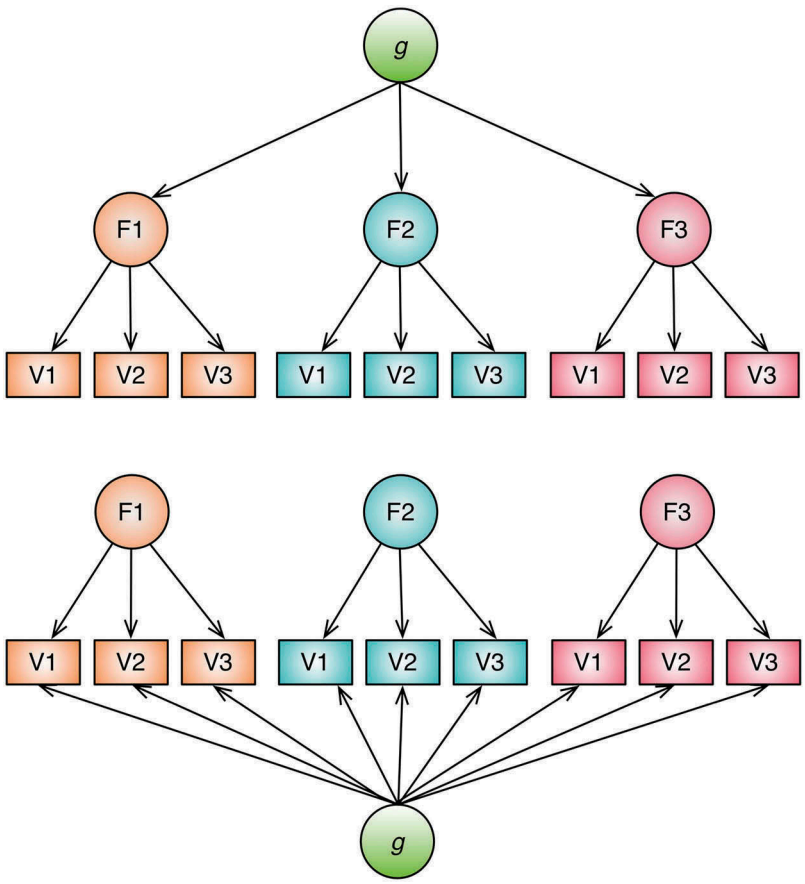

Figure 2. Conceptual illustration of higher-order (top panel) versus bifactor (bottom panel) models.

likelihood (ML) estimations are typically applied with multivariate continuous data like WISC-V^{Spain} scores. Estimation method can affect model fit and parameter estimates, and are sensitive to sample size and multivariate normality (Lei & Wu, 2012). 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." The effect of WLS on WISC-V^{Spain} estimates is unknown but should be investigated.

Fourth, the five-factor model preferred by Wechsler (2015b) abandoned the parsimony of simple structure (Thurstone, 1947) by allowing multiple cross-loadings of the Arithmetic subtest (see Figure 1). 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 et al., 2015). In fact, simple structure is implied by the scoring structure of the WISC-V^{Spain} where composite scores are created from unit-weighted sums of subtest scores. Thus, the preferred five-factor model is discrepant from the actual scoring structure of the WISC-V^{Spain}.

Fifth, Wechsler (2015b) relied on chi-square difference tests of nested models that are sensitive to trivial differences with large samples. Millsap (2007, p. 878) 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." Wechsler (2015b) reported nine 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. Philosophers of science have repeatedly warned about the dangers of null hypothesis significance testing (Haig, 2017) and the validity of chi-square difference tests has been contested (Yuan & Chan, 2016). Consequently, the differences in global fit relied on by Wechsler (2015b) to select preferred models might reflect only trivial differences between models. For example, Wechsler's (2015b) models 5c and 5d exhibited identical fit indices but their $\Delta\chi^2$ was statistically significant.

A sixth concern is that global model fit, by itself, does not guarantee model veracity (Bowen & Guo, 2012; Brown, 2015; DiStefano & Hess, 2005; Kline, 2016; MacCallum & Austin, 2000; Stromeyer et al., 2015; Widaman, 2012). Even with good global fit, relationships among variables might be weak, parameter estimates might not be statistically or practically significant, the latent variables might not account for meaningful variance in the indicators, or parameter values might be out of range. Wechsler (2015b) did not report the statistical significance of parameter estimates but a review of Figure 1, the publisher preferred structural model for the WISC-V^{Spain}, reveals several areas of local stress. For example, standardized loadings of .21 and .28 for the Arithmetic subtest and a standardized path coefficient of .97 between the higher-order general intelligence factor and the first-order Fluid Reasoning (FR) factor. This almost-perfect relationship constitutes a threat to discriminant validity (Brown, 2015; Kline, 2016) and indicates that the *g* and FR factors were empirically redundant (Le et al., 2010).

Finally, Wechsler (2015b) did not report the proportions of variance accounted for by general and group factors, nor the communality of measured variables. These metrics speak to the relationships of measured and latent variables and are important for accurate interpretation of common factors (Brown, 2015; Gignac & Kretzschmar, 2017; MacCallum & Austin, 2000). Further, they allow the computation of model-based reliability estimates that replace the classical test theory hypothesis of true and error variance with the factor analytic conceptualization of common and unique variance while simultaneously making fewer and more realistic assumptions than coefficient alpha (Gignac, 2015; McDonald, 1999; Reise, 2012; Rodriguez, Reise, & Haviland, 2016; Watkins, 2017; Zinbarg, Revelle, Yovel, & Li, 2005). Consequently, these estimates provide "a better estimate for the composite score and thus should be used" (Chen, Hayes, Carver, Laurenceau, & Zhang, 2012, p. 228).

In summary, the evidence provided by Wechsler (2015b) to support the factorial validity of the WISC-V^{Spain} is open to question. However, competent psychological assessment demands strong supportive evidence of reliability and validity before any test, including the WISC-V^{Spain}, can be used to make high-stakes decisions (AERA, APA, & National Council on Measurement in Education [NCME], 2014; British Psychological Society, 2007; Evers et al., 2013; ITC, 2013; Krishnamurthy et al., 2004). Accordingly, the factor structure of the WISC-V^{Spain} was independently analyzed to identify an appropriate scoring structure for all 15 subtests as well as for the 10 subtests likely to be used in clinical practice.

Method

Participants

Participants were the WISC-V^{Spain} standardization sample of 1,008 children aged 6 years to 16 years of

age, stratified by age, sex, parent education level, geographic region (North, South, East, and West), and geographic type (i.e., rural, suburban, and urban). The sample appeared to be representative of Spanish children (see Wechsler, 2015b for full details).

Instrument

The WISC-V^{Spain} is a norm-referenced, individually administered intelligence battery appropriate for children aged 6 through 16 years. According to the WISC-V^{Spain} 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^{Spain} scores.

As illustrated in Figure 1, the WISC-V^{Spain} contains 10 primary and 5 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 5 secondary subtests are proposed to load onto the same factors as the primary subtests: Information (IN) and Comprehension (CO) on the VC factor; Arithmetic (AR) on the FR factor; Letter-Number Sequencing (LN) on the WM factor; and Cancellation (CA) on the PS factor.

Wechsler (2015b) reported that the average split-half reliability of the FSIQ for the total standardization sample was .95, whereas the average reliability of factor index scores ranged from .88 (PSI) to .93 (FRI) and the average reliability of subtest scores ranged from .74 (CO) to FW (.93). Short-term test–retest reliability was .89 for the FSIQ, ranged from .74 (FRI) to .87 (VSI) for the factor index scores, and ranged from .67 (FW) to .84 (CA) for subtests. Concurrent validity was supported by a comparison of WISC-V^{Spain} scores to other cognitive and achievement tests. Convergent and discriminant validity was supported by studies of WISC-V^{Spain} scores among clinical groups. Factorial validity evidence was presented via a series of CFA with the final structural model adhering to a CHC framework (see Figure 1).

Analyses

Correlations, means, and standard deviations of the 15 WISC-V^{Spain} primary and secondary subtests for the total standardization sample were extracted from Wechsler (2015b). All CFA were conducted with Mplus 8.0 (Muthén & Muthén, 2017) 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).

Analyses

The evaluated models were taken from Wechsler (2015b) and are detailed in Table 1. Wechsler (2015b) only included higher-order models with general intelligence at the second level (see Figure 1 and the top panel of Figure 2 for examples). Analyses with the 10 primary

Table 1. Alternative structural models for the WISC-V^{Spain} with 15 primary and secondary subtests.

	1	2	3	4a	4b	4c	4d	5a	5b	5c	5d	5e
SI*	F1	F1	F1	F1	F1	F1	F1	F1	F1	F1	F1	F1
VO*	F1	F1	F1	F1	F1	F1	F1	F1	F1	F1	F1	F1
IN	F1	F1	F1	F1	F1	F1	F1	F1	F1	F1	F1	F1
CO	F1	F1	F1	F1	F1	F1	F1	F1	F1	F1	F1	F1
BD*	F1	F2	F2	F2	F2	F2	F2	F2	F2	F2	F2	F2
VP	F1	F2	F2	F2	F2	F2	F2	F2	F2	F2	F2	F2
MR*	F1	F2	F2	F2	F3	F2	F2	F3	F3	F3	F3	F3
FW*	F1	F2	F2	F2	F3	F2	F2	F3	F3	F3	F3	F3
AR	F1	F1	F1	F3	F3	F2-F3	F1-F3	F4	F3	F3-F4	F1-F4	F1-F3-F4
DS*	F1	F1	F1	F3	F3	F3	F3	F4	F4	F4	F4	F4
PS	F1	F2	F2	F3	F3	F3	F3	F4	F4	F4	F4	F4
LN	F1	F1	F1	F3	F3	F3	F3	F4	F4	F4	F4	F4
CD*	F1	F2	F3	F4	F4	F4	F4	F5	F5	F5	F5	F5
SS	F1	F2	F3	F4	F4	F4	F4	F5	F5	F5	F5	F5
CA	F1	F2	F3	F4	F4	F4	F4	F5	F5	F5	F5	F5

Note. F1–F5 indicate the factor on which each subtest loads. SI = Similarities, VO = Vocabulary, IN = Information, CO = Comprehension, BD = Block Design, VP = Visual Puzzles, MR = Matrix Reasoning, FW = Figure Weights, AR = Arithmetic, DS = Digit Span, PS = Picture Span, LN = Letter-Number Sequencing, CD = Coding, SS = Symbol Search, and CA = Cancellation. Models 2-5c are higher-order, as reported by Wechsler (2015b). The 10 primary subtests are in bold and subtests that contribute to the Full Scale IQ score are marked with an *.

subtests used to create factor index scores were conducted in addition to analyses with all 15 primary and secondary subtests. Bifactor versions of simple structure models were also included to allow a comparison of alternative models not tested by Wechsler (2015b).

Global model fit was evaluated with the chi-square likelihood ratio, comparative fit index (CFA), Tucker-Lewis index (TLI), 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 (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 RMSEA ≤ .06. The AIC was used to compare the global fit of alternative models with the lowest AIC value indicating the model most likely to generalize (Akaike, 1987). Meaningful differences between well-fitting models was also evaluated using ΔCFI/TLI > .01, ΔRMSEA > .015 (Chen, 2007; Cheung & Rensvold, 2002; Gignac, 2007), and ΔAIC ≥ 10 (Anderson, 2008). 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 also examined to ensure that they made statistical and substantive sense (Brown, 2015).

Model-based reliability

The most popular model-based reliability estimates are omega coefficients (Gignac, 2015; McDonald, 1999; Reise, 2012; Rodriguez et al., 2016; Watkins, 2017; Zinbarg et al., 2005). The most general omega coefficient is omega total (ω), which is an "estimate of the proportion of variance in the observed total score attributable to all 'modeled' sources of common variance" (Rodriguez et al., 2016, p. 140). High ω values indicate a highly reliable *multidimensional* total score. Another omega coefficient, called omega subscale (ω_s), can be computed for each unit-weighted subscale score. ω_s indexes the proportion of variance in each unit-weighted subscale score attributable to a blend of general and group factor variance. High ω_s values indicate a highly reliable *multidimensional* group factor score. Like coefficient alpha, ω and ω_s both reflect the systematic variance attributable to multiple common factors and neither can distinguish between the precision of the general factor versus the precision of the group factor. There is no universally accepted guideline for coefficient alpha values sufficient for high-stakes decisions about individuals, but values ≥ .90 are commonly recommended and "scores with values below .90 should not be interpreted" (Kranzler & Floyd, 2013, p. 71). Given their conceptual similarity, omega coefficients should meet the same standard as alpha coefficients.

Another omega variant allows a distinction between general and group factor variance because it indexes variance attributable to a single factor independent of all other factors and is, therefore, a measure 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 represents the strength of the general factor. When applied to group factors, omega hierarchical subscale (ω_{hs}) indexes 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 (Reise, 2012). There is also no empirically based guideline for acceptable levels of omega hierarchical coefficients for clinical decisions about individuals, but it has been suggested that they should, at a minimum, exceed .50 although .75 would be preferred (Reise, 2012). Given this standard, "the meaningful interpretation of index scores is arguably impossible" when omega hierarchical coefficients drop below .50 (Gignac & Watkins, 2013, p. 658).

The *H* coefficient of Hancock and Mueller (2001) provides another perspective on construct reliability. Whereas omega hierarchical represents the correlation between a factor and a unit-weighted composite score, *H* is the correlation between a factor and an optimally weighted composite score and represents how well a latent variable is represented by its indicators. When *H* is low, the latent variable is not very well defined by its indicators and will tend to be unstable across studies (Rodriguez et al., 2016). Hancock and Mueller (2001) suggested criterion values of *H* ≥ .70 for sufficient certainty regarding the relations among constructs.

Results

15 Primary and secondary subtests

Global fit measures for all tested models are reported in Table 2. Models with fewer than four group factors

Table 2. Confirmatory factor analysis fit statistics for the WISC-V^{Spain} total standardization sample (N = 1,008).

Model	χ^2	df	CFI	ΔCFI	TLI	ΔTLI	RMSEA	90% CI RMSEA	AIC	ΔAIC
15 Subtests										
1	1155.7	90	.825	.159	.796	.182	.108	.103–.114	70896.9	953.2
2	995.6	89	.851	.133	.824	.154	.101	.095–.106	70738.8	795.1
3	647.6	87	.908	.076	.889	.089	.080	.074–.086	70394.7	451.0
4a	315.5	86	.962	.022	.954	.024	.051	.045–.058	70064.6	120.9
4a bifactor	174.8	75	**.984**	**0**	.977	.001	**.036**	**.029–.043**	69945.9	2.2
4b	386.3	86	.951	.033	.940	.038	.059	.053–.065	70135.5	191.8
4b bifactor	204.9	76	.979	.005	.971	.007	.041	.034–.048	69974.1	30.4
4c	261.2	85	.971	.013	.964	.014	.045	.039–.052	70012.3	68.6
4d	240.3	84	.974	.010	.968	.010	.043	.037–.049	69993.5	49.8
5a	297.4	85	.965	.019	.957	.021	.050	.044–.056	70048.5	104.8
5a bifactor[b]	201.9	77	.979	.005	.972	.006	.040	.033–.047	69969.0	25.3
5b[a]	242.1	85	.974	.010	.968	.010	.043	.037–.049	69993.2	49.5
5b bifactor[b]	211.3	76	.978	.006	.969	.009	.042	.035–.049	69980.5	36.8
5b bifactor FR-VS[bc]	174.5	76	**.984**	**0**	**.978**	**0**	**.036**	**.029-.043**	**69943.7**	**0**
5c	227.9	84	.976	.008	.970	.008	.041	.035–.048	69981.1	37.4
5d	234.5	84	.975	.009	.969	.009	.042	.036–.049	69987.6	43.9
5e	218.2	83	.978	.006	.972	.006	.040	.034–.047	69973.3	29.6
5e bifactor	196.1	74	.980	.004	.972	.006	.040	.034-.047	69969.2	25.5
10 Subtests										
4a	122.9	31	.972	.015	.960	.019	.054	.044-.064	47507.2	46.3
4a bifactor[b]	70.7	28	**.987**	**0**	**.979**	**0**	**.039**	.028-.050	**47460.9**	**0**
5a[d]	89.5	30	.982	.005	.973	.006	.044	.034-.055	47475.7	14.8
5a bifactor[bd]	89.5	30	.982	.005	.973	.006	.044	.034-.055	47475.7	14.8

Note. CFI = Comparative Fit Index, TLI = Tucker–Lewis Index, RMSEA = root mean square error of approximation, AIC = Akaike's Information Criterion. Models correspond to those listed in Table 1 (plus bifactor variants of simple structure models that were added for this study). Each index of best-fitting model in bold. Indices not meaningfully different (ΔCFI and ΔTLI < .01, ΔRMSEA > .015, ΔAIC ≤ 10) from best fit shaded.

[a]Improper solution. Negative error variance estimate for FR factor.

[b]Models with only two indicators for group factors were constrained to equality for identification.

[c]Identical to bifactor model 5b except the Fluid Reasoning and Visual Spatial group factors were allowed to correlate as suggested by Reynolds and Keith (2017).

[d]Not statistically distinguishable due to the constraints needed to identify five factors with only 10 indicators

failed to achieve good global fit, whereas models with four and five group factors generally attained good global fit. All models with good fit were also inspected for size of parameters, statistical significance of parameters, and interpretability. All higher-order models with a CHC structure were marked by FR loadings on the general intelligence factor at such high levels (.96 to 1.03) as to indicate that those factors were empirically redundant (Bowen & Guo, 2012; Le et al., 2010). Although these models exhibited good global fit, they were invalidated by parameters that were statistically or substantively improper.

The initial bifactor version of model 5b was improper, exhibiting a negative variance estimate for the FR factor. However, that model converged appropriately when the FR and VS factors were allowed to correlate (as per Reynolds & Keith, 2017) and exhibited the best global fit (see right panel of Figure 3). The bifactor version of the traditional Wechsler structure (model 4a) was the second-best-fitting model (see left panel of Figure 3). The FR and VS factors merged in this model, although the residualized loadings of the MR and FW subtests were smaller than those of the BD and VP subtests. These two models were not statistically nor meaningfully different. Additionally, all parameters were statistically significant, none were out of range, and all were substantively meaningful in both models.

Given these results, it appears that a four-factor Wechsler model and a modified five-factor CHC model exhibited equivalent fit to the data. However, the CHC model was marked by low discriminant validity with concomitant interpretational confounding (Stromeyer et al., 2015).

The bifactor version of the traditional Wechsler model (e.g., 4a) was selected for variance decomposition and computation of model-based reliability coefficients because it afforded greater parsimony through simple structure and fewer factors. Sources of variance from that model are presented in Table 3. The general factor accounted for 35.3% of the total variance and 66.5% of the common variance. Only the PS group factor accounted for more than 5% of the total variance. In fact, the general factor accounted for about twice the total and common variance of all group factors combined. All together, the general and group factors accounted for 53% of the total variance leaving 47% due to specific variance and error.

Only the AR subtest was a good measure of *g* whereas three subtests (CD, SS, and CA) were poor measures of *g* (Kaufman, 1994). Communality of the CA subtest was very low, only 26% of its variance was explained by the general and PS factors. Thus, 74% of its variance was specific and error variance. The explained common variance (ECV; Rodriguez et al., 2016) for a subtest is the ratio of variance explained by a general factor divided by the variance explained by both general and group factors (see Table 3). Using that metric, more than 90% of the variance of the MR, FW, and AR subtests was explained by the general factor.

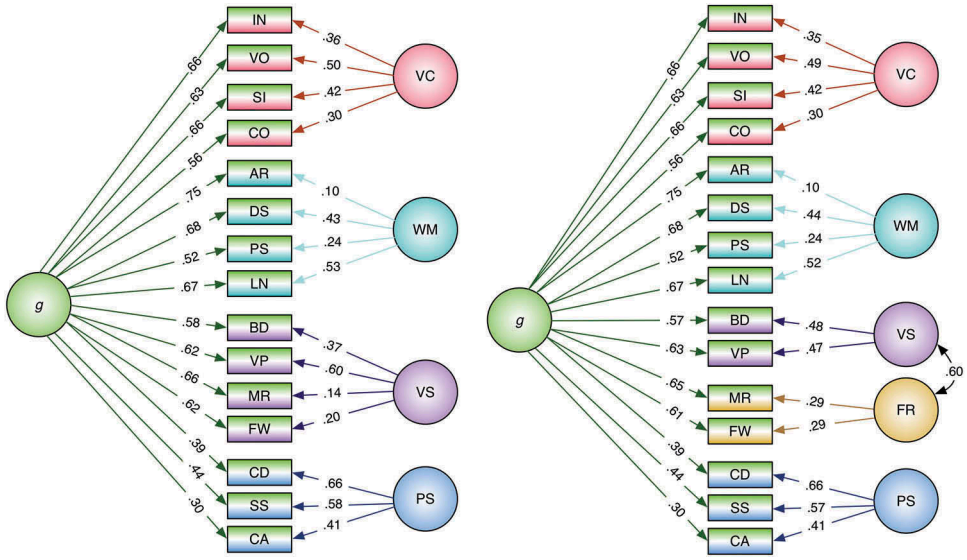

Figure 3. Best-fitting models among the 15 WISC-V^{Spain} primary and secondary subtests. Left panel is bifactor model 4a. Right panel is bifactor model 5b with correlated FR and VS factors.

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

Table 3. Sources of variance for WISC-V^{Spain} standardization sample (N = 1,008) 15 primary and secondary subtests 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	ECV
Similarities	*.66*	.436	.42	.176							.606	.394	.71
Vocabulary	*.63*	.397	.50	.250							.638	.362	.61
Information	*.66*	.436	.36	.130							.565	.435	.77
Comprehension	*.56*	.314	.30	.090							.404	.596	.78
Block Design	*.58*	.336			.37	.137					.472	.528	.71
Visual Puzzles	*.62*	.384			.60	.360					.741	.259	.52
Matrix Reasoning	*.66*	.436			.14	.020					.454	.546	.96
Figure Weights	*.62*	.384			.20	.040					.420	.580	.91
Arithmetic	***.75***	.563					.10	.010			.568	.432	.98
Digit Span	*.68*	.462					.43	.185			.647	.353	.71
Picture Span	*.52*	.270					.24	.058			.323	.677	.82
Letter–Number Sequencing	*.67*	.449					.53	.281			.724	.276	.62
Coding	.39	.152							.66	.436	.587	.413	.26
Symbol Search	.44	.194							.58	.336	.524	.476	.37
Cancellation	.30	.090							.41	.168	.261	.739	.35
Total Variance		35.3		4.3		3.7		3.6		6.3	53.2	46.8	
ECV		66.5		8.1		7.0		6.7		11.8			
H		.900		.443		.439		.408		.597			

Note. *b* = standardized loading of subtest on factor; b^2 = variance explained in the subtest; h^2 = communality; u^2 = uniqueness; and ECV = explained common variance, *H* is the correlation between a factor and an optimally weighted composite score (Hancock & Mueller, 2001). *g* loadings ≥ .70 are considered good (bold), from .50 to .69 are fair (italic), and < .50 are poor (Kaufman, 1994).

Model-based reliability coefficients are presented in Table 5. Only the FSIQ met the guideline of ≥ .90 for decisions about individuals, although the VC, WM, and VS factor scores were moderately reliable with ω_s coefficients of .81 to .83. 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} = .16 to .48), suggesting that "the apparent reliability of subscales judged by coefficient omega mostly is attributable to individual differences on the general factor" (Rodriguez et al., 2016, p. 142). For example, 83% of the variance of the unit-weighted VCI score was attributable to both general and VC factors, whereas only 24% of the variance of that same unit-weighted VCI score was uniquely attributable to the VC factor.

This situation would not be improved by utilizing optimally weighted composite scores as indicated by the *H* values in Table 3. The optimal composite of indicators explained 90% of the variability in the general intelligence construct but considerably less than 70% of the variability in the group factors. Thus, the group factors (VC, VS, WM, and PS) were not very well defined by their subtest indicators.

10 Primary subtests

Although there are 15 WISC-V^{Spain} subtests, only 10 are needed to produce its five factor scores. As a result, this 10-subtest scoring structure was also subjected to CFA. A four-factor model (4a) was created by combining the FR and VS subtests into a single factor. That bifactor model was statistically and practically superior in fit to its equivalent higher-order alternative (see Table 2).

Higher-order and bifactor models with five group factors were not statistically distinguishable due to the constraints needed to identify five factors with only 10 indicators (see Table 2). Nevertheless, bifactor model 4a was more likely to generalize than higher-order and bifactor five-factor models according to ΔAIC values. Thus, variance decomposition and model-based reliability coefficients for that four-factor bifactor model are presented in Tables 4 and 5, respectively. In this model, the general factor accounted for 34.8% of the total variance and 62.5% of the common variance, more than contributed by all group factors combined. The general and group factors accounted for 55.7% of the total variance leaving 44.3% due to specific variance and error. None of the subtests were good measures of *g* whereas two subtests (CD and SS) were poor measures of *g* (Kaufman, 1994). More than 95% of the variance of the MR and FW subtests was explained by the general factor.

Table 5. Omega reliability coefficients for WISC-V^{Spain} standardization sample (*N* = 1,008) from alternative models.

	Bifactor Wechsler 15 Subtests		10 Primary Subtests		7 FSIQ 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	**.92**	**.83**	**.90**	**.79**	.86	**.79**
Verbal Comprehension	.83	.24	.76	.27	.76	.26
Working Memory	.83	.16	.64	.15	.56	.10
Visual Spatial	.81	.18	.83	.16	.56	.18
Fluid Reasoning	–	–	–	–	.66	.08
Processing Speed	.71	.48	.71	**.50**	.32	.18

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 ~.90 for decisions about individuals (Kranzler & Floyd, 2013). At a minimum, omega hierarchical coefficients should exceed .50 although .75 would be preferred (Reise, 2012). Coefficients meeting minimum standards are in bold

Notably, the MR subtest had a statistically nonsignificant loading of .06 and the FW subtest had a practically nonsignificant loading of .14 on the Visual Spatial factor in that bifactor model, suggesting that general intelligence alone was responsible for variation in MR and FW subtest scores. That suggestion was reinforced by the results from the bifactor 5a model where neither the loading of MR nor FW was statistically significant.

Only the FSIQ was sufficiently reliable for decisions about individuals (ω = .90 and ω_h = .79). After controlling for the influence of the general factor, the five group factor scores were unreliable (ω_{hs} = .15 for WM to .50 for PS), each score providing "little information beyond that provided by the general factor estimate" (DeMars, 2013, p. 374). However, these reliability estimates are hypothetical for the FSIQ because the actual scoring structure of the WISC-V^{Spain} uses only 7

Table 4. Sources of variance for WISC-V^{Spain} standardization sample (*N* = 1,008) 10 primary subtests according to a bifactor Wechsler model.

	Generale		VC		VS		WM		PS				
Subtest	*b*	b^2	*b*	b^2	*b*	b^2	*b*	b^2	*b*	b^2	h^2	u^2	ECV
SI	*.6400*	.413	.47	.218							.632	.368	.655
VO	*.61*	.371	.47	.218							.589	.411	.630
MR	*.70*	.490			.06	.004					.494	.506	.993
FW	*.63*	.396			.14	.021					.416	.584	.950
BD	*.60*	.364			.27	.074					.438	.562	.831
VP	*.62*	.387			.75	.567					.954	.046	.406
DS	*.67*	.454					.33	.106			.561	.439	.810
PS	*.53*	.276					.33	.106			.382	.618	.722
CD	.39	.149							.62	.386	.535	.465	.279
SS	.43	.181							.62	.386	.566	.434	.319
ETV		34.8		4.4		6.7		2.1		7.7	55.7	44.3	
ECV		62.5		7.8		12.0		3.8		13.9			
H		.851		.358		.586		.192		.557			

Note. VC = Verbal Comprehension factor, FR = Fluid Reasoning factor, VS = Visual Spatial factor, WM = Working Memory factor, PS = Processing Speed factor, *b* = standardized loading of subtest on factor; b^2 = variance explained in the subtest, h^2 = communality, u^2 = uniqueness, SI = Similarities, VO = Vocabulary, IN = Information, CO = Comprehension, MR = Matrix Reasoning, FW = Figure Weights, AR = Arithmetic, DS = Digit Span, PS = Picture Span, LN = Letter-Number Sequencing, BD = Block Design, VP = Visual Puzzles, CD = Coding, SS = Symbol Search, CA = Cancellation, ETV = explained total variance, *H* is the correlation between a factor and an optimally weighted composite score (Hancock & Mueller, 2001), and ECV = explained common variance. *g* loadings ≥ .70 are considered good (bold), from .50 to .69 are fair (italic), and < .50 are poor (Kaufman, 1994)

subtests to compute the FSIQ, 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 for the FSIQ were .86 and .79, respectively. Thus, the shortened FSIQ was reduced in reliability but probably sufficiently precise for some individual decisions.

As with the 15-subtest model, this situation would not be improved by utilizing optimally weighted composite scores as indicated by the *H* values in Table 4. The optimal composite of indicators can explain 85% of the variability in the general intelligence construct, but less than 70% of the variability in the group factors. Thus, the group factors (VC, VS, WM, and PS) were not very well defined by their subtest indicators.

Discussion

Standardization sample data from the WISC-V^{Spain} were analyzed to investigate the construct validity of its scores. Although Wechsler (2015b) preferred a complex higher-order CHC model, the new FR factor in that model was problematic because it was empirically redundant with *g* and exhibited low reliability. 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^{Spain}. Alternatively, a bifactor CHC model with correlated FR and VS factors exhibited good fit but low discriminant validity with concomitant interpretational confounding (Stromeyer et al., 2015).

These results are not surprising given that previous Wechsler scales as well as other national versions of the WISC-V have exhibited similar bifactor structures (Canivez et al., 2016, 2016, 2017; Cucina & Byle, 2017; Dombrowski et al., 2017; Gignac & Watkins, 2013; Gomez, Vance, & Watson, 2017; Gustafsson & Undheim, 1996; Lecerf & Canivez, 2017; Styck & Watkins, 2016; Watkins, 2006; Watkins, Canivez, James, James, & Good, 2013; Watkins et al., 2017). For example, Canivez et al. (2017) applied CFA to scores from the U.S. WISC-V normative sample and found that a bifactor model with four group factors (where the FR and VS dimensions collapsed into a single factor) was most likely to generalize. As with the WISC-V^{Spain}, subtests did not saliently load on the FR group factor after the influence of general intelligence was taken into account, general intelligence accounted for more common and total variance than the group factors combined, and none of the group factor scores was sufficiently reliable for confident interpretation.

The merits of bifactor versus higher-order models have 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.

Murray and Johnson (2013) suggested that both bifactor or higher-order models 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, Nagy, & Wilhelm, 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^{Spain} 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, 2015b), 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 demonstrated that reliable variance of all WISC-V^{Spain} factor index scores was primarily due to the general factor, *not* the group factor (see Table 5). Contrary to Wechsler (2014b), the WISC-V index scores are *not* "reliable and valid measures of the primary cognitive constructs they intend to represent" (p. 149). Rather, around 45% of the total variance of WISC-V^{Spain} scores was due to error and specific variance and none of the group factor scores was sufficiently reliable for confident interpretation.

Although the publisher proposed a five-factor structure for the WISC-V^{Spain}, this study found that a traditional Wechsler four-factor structure was more appropriate when best-practice CFA methods were applied. The publisher also proposed interpretation of WISC-V^{Spain} scores at total (FSIQ), primary (VCI, VSI, FRI, WMI, PSI), and subtest levels. However, the interpretation chapter in Wechsler (2015b) emphasized the importance of factor index scores and only devoted one paragraph to the FSIQ. As with prior Wechsler

manuals, "contradictory findings available in the literature" (Braden & Niebling, 2012, p. 744) were not reported and there was no recognition that interpretation of group factor scores confounds variance contributed by group and general factors (Braden, 2013).

Psychologists should remember that "supporters of new methods often base their advocacy on justifications that have not been thoroughly vetted" (Stromeyer et al., 2015, p. 492) and that test users are ultimately responsible for "appropriate test use and interpretation" (AERA, APA, & NCME, 2014, p. 141) by ensuring that "there is sufficient validity and reliability information to interpret the test's scores in the way suggested by the test developer and publisher" (Beaujean, 2015, p. 53). International test standards and professional ethical standards also require that psychologists be aware of the available evidence regarding reliability and validity of test scores (APA, 2002; British Psychological Society, 2007; ITC, 2013). To that end, this study demonstrated that psychologists can be reasonably confident in using the WISC-V^{Spain} FSIQ score for clinical decisions but cannot expect the factor index scores to be sufficiently reliable for decisions about individuals because those scores represent a blend of general *and* group abilities as well as error and contribute little information beyond that provided by the general factor (Beaujean, Parkin, & Parker, 2014; Canivez, 2016; Cucina & Howardson, 2017). Thus, the present study did not support the interpretation methods advocated by the publisher (Wechsler, 2015b) nor those recommended by popular textbook authors (Sattler, Dumond, & Coalson, 2016).

Interpretation of factor index scores should also be informed by external validity evidence (AERA, APA, & NCME, 2014; Hummel, 1998; Kranzler & Floyd, 2013). DeMars (2013) predicted that differential validity would be impaired by scores with low reliability. 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 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, Kranzler, & Floyd, 2016; Canivez, 2013; Canivez, Watkins, James, Good, & James, 2014; Glutting, Watkins, Konold, & McDermott, 2006). Additionally, 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 decisions about individuals.

About the authors

Dr. Javier Fenollar-Cortés is Associate Professor of the Department of Developmental and Educational Psychology at University of Murcia (Spain). His research interests are in clinical assessment, school psychology and ADHD. He also works in collaboration with the Centre of Child Development and Early Attention HIGEA.

Dr. Marley W. Watkins is Professor and Chairman of the Department of Educational Psychology at Baylor University in Waco, TX. His research interests include the study of individual differences.

ORCID

Marley W. Watkins http://orcid.org/0000-0001-6352-7174

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*, 317–332. doi:10.1007/BF02294359

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.

American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist*, *57*, 1060–1073. doi:10.1037/a0020168

Anderson, D. R. (2008). *Model based inference in the life sciences: A primer on evidence*. New York, NY: Springer.

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. 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. (2013). Psychological assessment in school settings. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of psychology: Assessment psychology* (Vol. 10, 2nd ed., pp. 291–314). Hoboken, NJ: Wiley.

Braden, J. P., & Niebling, B. C. (2012). Using the joint test standards to evaluate the validity evidence for intelligence tests. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 739–757). New York, NY: Guilford.

Braden, J. P., & Shaw, S. R. (2009). Intervention validity of cognitive assessment: Knowns, unknowables, and unknowns. *Assessment for Effective Intervention*, *34*, 106–115. doi:10.1177/1534508407313013

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

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.

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. (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., 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*, *29*, 458–472. 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. doi:10.1037/h0088800

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

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

Chen, H., Zhang, O., Raiford, S. E., Zhu, J., & Weiss, L. G. (2015). Factor invariance between genders on the Wechsler Intelligence Scale for Children-Fifth Edition.. *Personality and Individual Differences*, *86*, 1–5. doi:10.1016/j.paid.2015.05.020

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

Cortina, J. M., Green, J. P., Keeler, K. R., & Vandenberg, R. J. (2017). Degrees of freedom in SEM: Are we testing the models that we claim to test? *Organizational Research Methods*, *20*, 350–378. doi:10.1177/1094428116676345

Cucina, J., & Byle, K. (2017). The bifactor model fits better than the higher-order model in more than 90% of comparisons for mental abilities test batteries. *Journal of Intelligence*, *5*(3), 27. doi:10.3390/jintelligence5030027

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

Daniel, M. H. (2007). "Scatter" and the construct validity of FSIQ: Comment on Fiorello et al. (2007). *Applied Neuropsychology*, *14*, 291–295. doi:10.1080/09084280701719401

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. doi:10.1177/073428290502300303

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

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., McCormick, C. M., Hawley, L. R., Muñiz, J., Balboni, G., Bartram, D., & Zhang, J. (2017). Testing practices and attitudes toward tests and testing: An international survey. *International Journal of Testing*, *17*, 158–190. doi:10.1080/15305058.2016.1216434

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. (2015). Estimating the strength of a general factor: Coefficient omega hierarchical. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *8*, 434–438. doi:10.1017/iop.2015.59

Gignac, G. E., & Kretzschmar, A. (2017). Evaluating dimensional distinctness with correlated-factor models:

Limitations and suggestions. *Intelligence*, *62*, 138–147. doi:10.1016/j.intell.2017.04.001

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. doi:10.1177/00224669060400020101

Gomez, R., Vance, A., & Watson, S. (2017). Bifactor model of WISC-IV: Applicability and measurement invariance in low and normal IQ groups. *Psychological Assessment*, *29*, 902–912. 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.

Haig, B. D. (2017). Tests of statistical significance made sound. *Educational and Psychological Measurement*, *77*, 489–506. doi:10.1177/0013164416667981

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.

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*(2), 93-114.*10.1207/S15327574IJT0102_1

International Test Commission. (2013). *International guidelines for test use (Version 1.2)*. Retrieved from https://www.intestcom.org

International Test Commission. (2017). *The ITC Guidelines for Translating and Adapting Tests (Second edition)*. Retrieved from https://www.intestcom.org

International Test Commission. (2016). *The ITC guidelines for translating and adapting tests* (2nd ed.). Retrieved from https://www.intestcom.org

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. (2013). *Assessing intelligence in children and adolescents: A practical guide*. New York, NY: Guilford.

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

Lecerf, T., & Canivez, G. L. (2017, December 28). Complementary exploratory and confirmatory factor analyses of the French WISC–V: Analyses based on the standardization sample. In *Psychological Assessment*. Advance online publication. http://dx.doi.org/10.1037/pas0000526

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.

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. doi:10.1146/annurev.psych.51.1.201

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: Lawrence Erlbaum.

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.

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? A comparison 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. (2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

Oakland, T., Douglas, S., & Kane, H. (2016). Top ten standardized tests used internationally with children and youth by school psychologists in 64 countries: A 24-year follow-up study. *Journal of Psychoeducational Assessment*, *34*, 166–176. doi:10.1177/0734282915595303

Piotrowski, C. (2017). Neuropsychological testing in professional psychology specialties: Summary findings of 36 studies (1990-2016) in applied settings. *Journal of the Indian Academy of Applied Psychology*, *43*, 135–145.

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

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. doi:10.1007/BF03340943

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). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*, 137–150. doi:10.1037/met0000045

Sattler, J. M., Dumond, R., & Coalson, D. L. (2016). *Assessment of children: WISC-V and WPPSI-IV*. San Diego, CA: Sattler.

Scheiber, C. (2016). Is the Cattell-Horn-Carroll-based factor structure of the Wechsler Intelligence Scale for Children–Fifth Edition (WISC-V) construct invariant for a representative sample of African-American, Hispanic, and Caucasian male and female students ages 6 to 16 years? *Journal of Pediatric Neuropsychology*, *2*, 79–88. doi:10.1007/s40817-016-0019-7

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

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

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

Watkins, M. W., 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., Dombrowski, S. C., & Canivez, G. L. (2017). Reliability and factorial validity of the Canadian Wechsler Intelligence Scale for Children–fifth edition. *International Journal of School & Educational Psychology*. Advance online publication. doi:10.1080/21683603.2017.1342580

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. doi:10.1080/09084280701280353

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

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

Wechsler, D. (2015a). *Escala de inteligencia de Wechsler para niños-V*. Madrid, Spain: Pearson Educación.

Wechsler, D. (2015b). *Escala de inteligencia de Wechsler para niños-V. Manual técnico y de interpretación*. Madrid, Spain: Pearson Educación.

Wechsler, D. (2016a). *Wechsler Intelligence Scale for Children–Fifth UK Edition*. London, UK: Harcourt Assessment.

Wechsler, D. (2016b). *Wechsler Intelligence Scale for Children, Fifth Edition: Australian and New Zealand Standardized Edition*. Sydney, Australia: Pearson.

Wechsler, D. (2016c). *Wechsler Intelligence Scale for Children and Adolescents–Fifth Edition: Adaptation Française*. Paris, France: ECPA.

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.

Wright, C. V., Beattie, S. G., Brabender, V. M., Smith, B. L., Galper, D. I., Church, A. S., & Bufka, L. F. (2017). Assessment practices of professional psychologists: Results of a national survey. *Professional Psychology: Research and Practice*, *48*, 73–78. doi:10.1037/pro0000086

Yuan, K.-H., & Chan, W. (2016). Measuring invariance via multigroup SEM: Issues and solutions with chi-square-difference tests. *Psychological Methods*, *21*, 405–426. doi:10.1037/met0000080

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