Multivariate Behavioral Research, 48:639–662, 2013

ISSN: 0027-3171 print/1532-7906 online
DOI: 10.1080/00273171.2013.804398

Routledge
Taylor & Francis Group

Bifactor Modeling and the Estimation of Model-Based Reliability in the WAIS-IV

Gilles E. Gignac
University of Western Australia

Marley W. Watkins
Baylor University

Previous confirmatory factor analytic research that has examined the factor structure of the Wechsler Adult Intelligence Scale–Fourth Edition (WAIS-IV) has endorsed either higher order models or oblique factor models that tend to amalgamate both general factor and index factor sources of systematic variance. An alternative model that has not yet been examined for the WAIS-IV is the bifactor model. Bifactor models allow all subtests to load onto both the general factor and their respective index factor directly. Bifactor models are also particularly amenable to the estimation of model-based reliabilities for both global composite scores (ω_h) and subscale/index scores (ω_s). Based on the WAIS-IV normative sample correlation matrices, a bifactor model that did not include any index factor cross loadings or correlated residuals was found to be better fitting than the conventional higher order and oblique factor models. Although the ω_h estimate associated with the full scale intelligence quotient (FSIQ) scores was respectably high (.86), the ω_s estimates associated with the WAIS-IV index scores were very low (.13 to .47). The results are interpreted in the context of the benefits of a bifactor modeling approach. Additionally, in light of the very low levels of unique internal consistency reliabilities associated with the index scores, it is contended that clinical index score interpretations are probably not justifiable.

Correspondence concerning this article should be addressed to Gilles E. Gignac, School of Psychology, University of Western Australia, 35 Stirling Highway, Crawley, Western Australia, 6009, Australia. E-mail: gilles.gignac@uwa.edu.au

Since the publication of the Wechsler Adult Intelligence Scale–Fourth Edition (WAIS-IV; Wechsler, 2008a), several studies have sought to extend the confirmatory factor analyses reported in the WAIS-IV technical manual (Wechsler, 2008b). Based on a series of competing models (i.e., higher order, oblique, Cattell-Horn-Carroll [CHC]) some convergence on a CHC interpretation of the WAIS-IV intersubtest covariation appears to be warranted (Benson, Hulac, & Kranzler, 2010; Ward, Bergman, & Hebert, 2011). However, one model that has yet to be tested for the WAIS-IV is the bifactor model (aka nested factor model, direct hierarchical model), which has been found to be a superior fitting model when tested on previous editions of the Wechsler scales (Gignac, 2005, 2006a). Furthermore, model-based estimates of reliability can be applied insightfully to bifactor models as they represent the amount of unique internal consistency reliability associated with both the general composite scores (e.g., full scale intelligence quotient [FSIQ]) and the narrower composite scores (e.g., index scores). Thus, the purpose of this article is to test the plausibility of a WAIS-IV index bifactor model as well as estimate the model-based reliabilities associated with the FSIQ and index composite scores.

PAST EMPIRICAL RESEARCH

The WAIS-IV consists of 15 subtests (10 core and 5 supplemental) designed to measure four positively intercorrelated indices: Verbal Comprehension (VC), Perceptual Reasoning (PR), Working Memory (WM), and Processing Speed (PS). As the four indices are positively intercorrelated, they can be used to form total scale scores (FSIQ; Wechsler, 2008b). To evaluate the factorial validity associated with the WAIS-IV, Wechsler (2008b) tested a series of competing models, including a higher order model (one second-order general factor and four first-order index factors; see Rindskopf & Rose, 1988, for a full description of a higher order model) and a corresponding oblique four-factor model (Wechsler, 2008b). The higher order model was endorsed by Wechsler (2008b), but it allowed two subtests to have interindex cross loadings (Arithmetic and Figure Weights) and also included correlated residual error terms between the Digit Span and Letter-Number Sequencing subtests. Thus, it may be suggested that the endorsed model was not associated with as simple a structure as would be desired (Bowen & Guo, 2012).

Benson et al. (2010) performed a series of confirmatory factor analyses (CFA) on portions of the WAIS-IV normative sample to evaluate the CHC (McGrew, 1997) as an alternative to the WAIS-IV index model. The main distinctions between the WAIS-IV index model and the CHC model tested in Benson et al. is that the PR index factor was split into two factors: Visual Processing (G*v*) and Fluid Reasoning (Gf). Additionally, the Arithmetic subtest was specified to load onto the Gf factor rather than a memory factor as per the WAIS-IV index model (Wechsler, 2008b). Based on Benson et al.'s calibration sample analyses ($N = 800$), evidence in favor of a CHC model interpretation was reported, as the higher order CHC model was associated with a lower Akaike Information Criterion (AIC; Akaike, 1973) value (382.70) in comparison with the higher order WAIS index model AIC value (499.52).

However, the CHC higher order model endorsed by Benson et al. (2010) is arguably of questionable interpretative value because one of the lower order factors appeared to be associated with a possible Heywood case. That is, the Gf factor was reported by Benson et al. to be associated with a higher order loading equal to 1.00. Thus, there is reason to question the plausibility of the higher order CHC model endorsed by Benson et al. as it may be overparameterized (Jöreskog & Sörbom, 1989).

Finally, Ward et al. (2011) examined via CFA the WAIS-IV from the perspective of the CHC model. However, they specified the Arithmetic subtest to have a cross loading on the Gf and Gc first-order factors. Additionally, Ward et al. applied an oblique factor modeling strategy, rather than a higher order modeling strategy, as they contended that an oblique factor model was associated with superior model fit in comparison with the Benson et al. (2010) endorsed higher order model. Additionally, an oblique factor modeling strategy overcomes the potential problem of a Heywood case as arguably observed within the Benson et al. endorsed higher order CHC model. However, a problem with the oblique factor model endorsed by Ward et al. is that it does not specify a general factor of intelligence, which is inconsistent with the overwhelming amount of empirical research in the area (Carroll, 1993) and "central to the Wechsler and other models of intelligence" (Wechsler, 2008b, p. 66). Additionally, the observation or incorporation of cross loadings within a factor model may be considered problematic and should be avoided if possible, as they complicate the interpretation of the corresponding composite scores (Bowen & Guo, 2012; Costello & Osborne, 2005).

One model that has yet to be tested for on the WAIS-IV is the bifactor model (Gustafsson & Balke, 1993; Holzinger & Swineford, 1937). Based on both the WAIS-R (Wechsler, 1981) and the WAIS-III (Wechsler, 1997) normative sample correlation matrices, Gignac (2005, 2006a) found the bifactor model to be associated with superior model fit in comparison with higher order and oblique factor models. However, Gignac (2005, 2006a) did not estimate the unique model-based reliabilities associated with the FSIQ scores or the index scores. Thus, we considered there to be compelling support for (a) testing a bifactor model on the WAIS-IV normative sample and (b) estimating the unique model-based internal consistency reliabilities associated with the FSIQ and index composites scores.

BIFACTOR MODEL

A bifactor model (aka nested factor model or direct hierarchical model) consists of one first-order general factor and one or more usually orthogonal, first-order factors nested within the general factor (Gustafsson & Balke, 1993; Holzinger & Swineford, 1937). In a typical bifactor model, each indicator is specified to load on the general factor directly and on one nested factor directly (see Model 3, Figure 1, of this article for a visual representation). However, cross loadings between nested factors can be specified as well. Additionally, there are theoretically interesting bifactor models that include indicators specified to load onto more than two orthogonal latent variables. For example, Gignac (2010) tested a bifactor model on a self-report emotional intelligence questionnaire whereby the negatively keyed items were specified to load onto the general factor, one domain-specific factor, and a negatively keyed item factor.

Bifactor model solutions can be estimated within both exploratory factor-analytic and confirmatory factor-analytic frameworks (Jennrich & Bentler, 2011; Reise, 2012). This article's focus is on bifactor models estimated within a confirmatory factor-analytic framework. It is important to note that a bifactor model solution is not necessarily the same as a Schmid-Leiman transformation (Schmid & Leiman, 1957) of a higher order model solution (Chen, West, & Sousa, 2006). When a generalized Schmid-Leiman transformation is applied to a higher order model solution (i.e., when proportionality constraints are imposed), meaningful interpretive differences can emerge between bifactor model solutions and Schmid-Leiman transformed higher order model solutions (e.g., Chen et al., 2006; Gignac, 2007a).

Although the bifactor model of intelligence within a confirmatory factor-analytic framework was introduced approximately 20 years ago (Gustafsson & Balke, 1993), it would probably be accurate to suggest that it has not yet been accepted to any appreciable degree (Gustafsson & Aberg-Bengtsson, 2010). For example, Keith (2005) found the bifactor model to be a superior fitting model for the Wechsler Intelligence Scale for Children–Fourth Edition (WISC-IV; Wechsler, 2003) normative sample but endorsed the higher order model on the grounds that the bifactor model "does not test an actual hierarchical model" (p. 594) and that it "is not consistent with any modern theoretical orientation" (p. 594).

However, as contended by Gignac (2008), the term *hierarchical model* used by early factor analysts such as Humphreys (1962) was used in the context of a completely first-order factor solution, not a higher order model solution. That is, the term *hierarchical model* is used to represent a model in which factors can be ranked by the number of subtests that define them. In the context of intelligence testing, the general factor has the greatest breadth, whereas the group-level factors have lesser levels of breadth. In this sense, the bifactor model

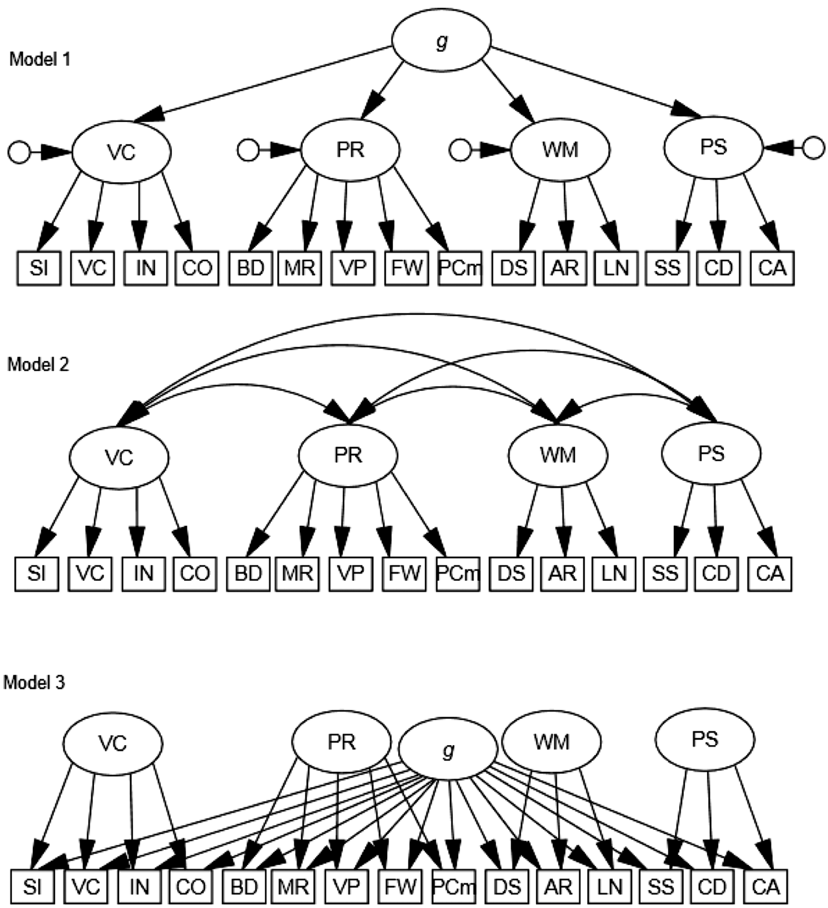

FIGURE 1 Series of competing models tested in this investigation; Model 1 = higher order WAIS-IV index factor model; Model 2 = oblique factor WAIS-IV index factor model; Model 3 = bifactor WAIS-IV index factor model; *g* = general factor; VC = Verbal Comprehension; PR = Perceptual Reasoning; WM = Working Memory; PS = Processing Speed; SI = Similarities; VC = Vocabulary; IN = Information; CO = Comprehension; BD = Block Design; MR = Matrix Reasoning; VP = Visual Puzzles; FW = Figure Weights; PCm = Picture Completion; DS = Digit Span; AR = Arithmetic; LN = Letter-Number Sequencing; SS = Symbol Search; CD = Coding; CA = Cancellation.

is in fact hierarchical. Higher order models, by contrast, emphasize differences between factors based on superordination (Gignac, 2008).

In addition to theoretical reservations, Keith (2005) contended that the observed superior fit associated with the WISC-IV bifactor model should probably be viewed as unusual. However, a nonnegligible amount of empirical research suggests that the observation of a superior fitting bifactor model is not unusual. In fact, a bifactor model has been shown to be associated with significant improvements in model fit over competing higher order models on both the WAIS-R and the WAIS-III normative sample intersubtest correlation matrices (Gignac, 2005, 2006a; Golay & Lecerf, 2011). For instance, Gignac (2005) estimated the model fit associated with a series of higher order and bifactor models for the WAIS-R normative sample. In one case, a conventional higher order model with one general factor and two lower order factors (verbal intelligence quotient [VIQ] and performance intelligence quotient [PIQ]) was tested against the corresponding bifactor model with one lower order general factor and two nested group-level factors (VIQ and PIQ). Gignac (2005) found that the bifactor model was associated with a Tucker-Lewis Index (TLI) = .973, which was considered a substantial improvement over the corresponding higher order model (TLI = .931).

The bifactor model results reported in Gignac (2005) also had practical implications. In particular, based on the bifactor model, the Arithmetic subtest was found not to contribute any statistically significant variance to the nested VIQ factor. By contrast, the corresponding higher order model failed to suggest that Arithmetic was not a valid indicator of VIQ. This was considered an important observation as the WAIS-R scoring guidelines specified the Arithmetic subtest as one of the six defining VIQ subtests (Wechsler, 1981).

In another investigation, Gignac (2006a) evaluated competing higher order and bifactor models based on the WAIS-III normative sample correlation matrices. Based on the total sample correlation matrix (N = 2,450), the CFA results largely favored a bifactor model interpretation. Specifically, the higher order index model endorsed by Wechsler (1997) was associated with a TLI = .959, whereas the corresponding bifactor model was associated with a TLI = .966. Perhaps more important, the bifactor model results did not suggest that the Arithmetic subtest should be considered a unique indicator of VIQ, as per Gignac (2005). Furthermore, another bifactor model tested by Gignac (2006a) found the Arithmetic subtest to be a negligible indicator (.16) of the WM index factor. Thus, the results associated with the bifactor modeling strategy had possible practical implications relevant to how the WAIS-III should be scored. That is, arguably, Arithmetic should probably not be included within a VIQ or WM index.

The results in support of a bifactor model interpretation of the Wechsler scales reported in Gignac (2005, 2006a) have been replicated on the French and Spanish versions of the WAIS-III (Golay & Lecerf, 2011; Molenaar, Dolan, & van der Maas, 2011). Additionally, the model fit superiority associated with WISC-IV observed by Keith (2005) has been replicated on a sample of 355 students referred for psychoeducational examination (Watkins, 2010). Furthermore, the Wechsler scales are not the only intelligence test batteries to have been shown to correspond more closely to a bifactor model. Others include the Multidimensional Aptitude Battery (MAB; Gignac, 2006b), the Berlin Intelligence Structure Test (BIS; Brunner & Süß, 2005), and the Swedish Enlistment Battery (Mårdberg & Carlstedt, 1998). Thus, in light of the nonnegligible amount of research endorsing a bifactor model of the Wechsler scales specifically, and intelligence test batteries more generally, it was considered useful to test the plausibility of a bifactor model for the WAIS-IV normative sample data.

MODEL-BASED INTERNAL CONSISTENCY RELIABILITY

Coefficient α is arguably the most common method used to estimate the internal consistency reliability of test scores (Peterson, 1994). Coefficient α is known as the ratio of true score variance to total variance (Lord & Novick, 1968). As per Cortina (1993), coefficient α may be formulated as

$$\alpha = \frac{k^2 * \overline{COV}}{\sum S^2, COV} \tag{1}$$

where k = number of indicators associated with the scale, $\overline{COV}$ = mean interindicator covariance, and $\sum S^2\ COV$ = the sum of the square variance/covariance matrix (i.e., composite score variance). Despite coefficient α's popularity, a number of papers have been published that highlight its limitations (e.g., Green & Yang, 2009; Schmitt, 1996; Sijtsma, 2009). Most pertinent to this article is the limitation that coefficient α will tend to conflate multiple sources of systematic variance when the data are associated with a multidimensional model (Lucke, 2005; Zinbarg, Revelle, Yovel, & Li, 2005).

An alternative approach to the estimation of internal consistency reliability is known as model-based internal consistency reliability (Bentler, 2009; Miller, 1995). There are now several forms of model-based internal consistency reliability, including ω (McDonald, 1985), ω_h (Zinbarg et al., 2005), and ω_s (Reise, Bonifay, & Haviland, 2012). In this article, the focus is on ω_h and ω_s.

Coefficient ω_h represents the unique internal consistency reliability associated with total scale composite scores (Zinbarg et al., 2005). Coefficient ω_h may be regarded as unique as it is applied to bifactor model solutions that represent the global factor as orthogonal to any nested factors. Thus, the true score variance associated with the global factor is independent of the true score variance associated with the nested factors. Consistent with Zinbarg et al. (2005) and Reise et al. (2012), ω_h may be formulated as

$$\omega_h = \tag{2}$$

$$\frac{\left.\sum_{i=1}^{k}\lambda_{g_i}\right)^2}{\left.\sum_{i=1}^{k}\lambda_{g_i}\right)^2 + \left.\sum_{i=1}^{s1}\lambda_{s1_i}\right)^2 + \left.\sum_{i=1}^{s2}\lambda_{s2_i}\right)^2 + \ldots + \left.\sum_{i=1}^{sp}\lambda_{sp_i}\right)^2 + \sum_{i=1}^{k}(1 = h_i^2)}$$

where λ_g corresponds to the general factor loadings; $\lambda_{s1}, \lambda_{s2} \ldots \lambda_{sp}$ correspond to the nested factor (or subscale) factor loadings; and $(1 - h^2)$ represents an indicator's unique variance (i.e., error). The denominator of Equation (2) corresponds to the total variance associated with the total scale's composite scores.

In contrast to ω_h, ω_s represents the amount of unique internal consistency reliability associated with subscale scores (Reise et al., 2012). Coefficient ω_s may be regarded as unique, as it is applied to bifactor model solutions that represent each nested factor as orthogonal to the general factor as well as (typically) orthogonal to any other nested factors. Thus, the true score variance associated with each nested factor is independent of the true score variance associated with any other nested factors and the global factor. Coefficient ω_s may be formulated as

$$\omega_s = \frac{\left.\sum_{i=1}^{s1}\lambda_{s1_i}\right)^2}{\left.\sum_{i=1}^{s1}\lambda_{g_i}\right)^2 + \left.\sum_{i=1}^{s1}\lambda_{s1_i}\right)^2 + \sum_{i=1}^{s1}(1 - h_{s1_i}^2)} \tag{3}$$

where all of the terms are defined as per Equation (2) with the exception that the elements are summed only for those indicators associated with a particular nested factor (in the aforementioned formula, s1 = subscale 1). The denominator of Equation (3) corresponds to the total variance associated with the subscale's composite scores.

Unlike coefficient α, model-based reliability estimates, such as the various ω coefficients, do not assume essential tau equivalence (Graham, 2006). Therefore, they tend not to be lower bound estimates of internal consistency reliability. Another advantage associated with model-based reliability is that it can be relatively easily applied to the bifactor model case, using either a formulation approach (Revelle, 2012; Zinbarg et al., 2005) or an approach based on the implied correlation between a phantom variable and its corresponding latent variable (Fan, 2003; Raykov, 1997). Furthermore, as the bifactor model specifies factors that are orthogonal to each other, each model-based reliability estimate may be considered the estimate of unique internal consistency reliability associated with each respective scale or subscale. That is, in the case of the Wechsler scales, conventional approaches (e.g., coefficient α) to the estimation of internal consistency reliability of index scores (e.g., Verbal Comprehension) fail to reflect the fact that an index's reliable variance is derived from two principal sources: g and the corresponding group-level factor.

To our knowledge, there are only three published studies that have estimated both ω_h and ω_s. In the first, Brunner and Süß (2005) estimated a bifactor model solution on a sample of 657 adults who completed the BIS (Jager, Süß, & Beauducel, 1997). The bifactor model consisted of one first-order general factor and seven nested group-level factors. Based on ω_h, the global index score (i.e., FSIQ) reliability was estimated to be .68, which was substantially lower than the original, conflated reliability estimate of .93 (i.e., coefficient α). Furthermore, the subscales were found to be associated with a mean ω_s equal to .31, which was also substantially lower than the mean of the original coefficient α reliabilities of .84. Thus, in both the total score and index score cases, the conventional approach to estimating internal consistency reliability via coefficient α, which imbues multiple sources of reliable variance, yielded substantial overestimates of internal consistency reliability.

In the second application of model-based ω_h and ω_s, Gignac, Palmer, and Stough (2007) investigated the factor structure of the self-report Toronto Alexithymia Scale (TAS-20; Bagby, Parker, & Taylor, 1994). Based on a sample of 363 respondents, Gignac, Palmer, et al. reported ω_s values of .56, .31, and .42 for the three TAS-20 subscales, which were all substantially lower than the original internal consistency reliability estimates (i.e., $> .70$). However, with respect to the total scale scores, Gignac, Palmer, et al. reported comparable ω_h and coefficient α values (.84, and .86, respectively). Thus, Gignac, Palmer, et al. found very low levels of internal consistency reliability unique to each subscale, as was observed in Brunner and Süß (2005). However, in contrast to Brunner and Süß, the TAS-20 total scores were found to be associated with an acceptable level of unique internal consistency reliability based on conventional standards (Nunnally & Bernstein, 1994). Finally, based on another bifactor model investigation of the TAS-20, ($N = 1{,}612$ college students), Reise et al. (2012) reported ω_h and ω_s results very similar to those reported in Gignac, Palmer, et al.

In light of the results reported in Brunner and Süß (2005); Gignac, Palmer, et al. (2007); and Reise et al. (2012), it was considered important to estimate model-based reliability (i.e., ω_h and ω_s) for one of the most well-known and commonly used psychological inventories: the WAIS-IV. It was hypothesized that the WAIS-IV subscales would be found to be associated with very low levels of unique internal consistency reliability as estimated via ω_s. Additionally, it was hypothesized that there would be some level of reduction in the internal consistency reliability estimation of the FSIQ scores as estimated via ω_h in comparison with the conventional application of coefficient α.

METHOD

Sample

All data analyses were based on the nine normative sample subtest correlation matrices published in the *WAIS-IV Technical and Interpretative Manual* (Tables A.1–A.9; Wechsler, 2008b). The WAIS-IV normative sample was obtained based on a stratified sampling strategy to reflect the U.S. census results relevant to gender, age, race/ethnicity, education, and geographic location (Wechsler, 2008b). As per Ward et al. (2011), the nine correlation matrices were combined to form four correlation matrices using a Fisher's z transformation and back-transformation procedure. Specifically, the ages 16–17 and 18–19 correlation matrices were combined to form a 16–19 correlation matrix ($N = 400$). The ages 20–24, 25–29, and 30–34 correlation matrices were combined to form a 20–34 correlation matrix ($N = 600$). The ages 35–44 and 45–54 correlation matrices were combined to form a 35–54 correlation matrix ($N = 400$). Finally, the ages 55–64 and 65–69 correlation matrices were combined to form a 55–69 correlation matrix ($N = 400$).

Data-Analytic Strategy

All confirmatory factor analyses (CFA) were performed with Amos 19.0 using maximum likelihood estimation (MLE). A series of four competing models were tested in this investigation (see Figure 1). Model 1 was the conventional higher order model with four first-order factors (VC, PR, WM, and PS) and one second-order general factor. Model 2 was the corresponding oblique factor model with the same four first-order factors that were specified in the aforementioned higher order model. Model 3 was the corresponding bifactor model, which allowed each subtest to load simultaneously onto the general factor directly and its corresponding index factor directly. Finally, Model 4 (not shown in Figure 1) was the CHC model endorsed by Benson et al. (2010) represented as a bifactor model. The main distinction between the CHC model and the WAIS-IV index factor model was that the PR factor was split into two factors: Spatial Visualization

(Gv) and Perceptual Reasoning (Gf). Additionally, the Arithmetic subtest was allowed to load onto both the Gf and Gsm factors.[1]

The higher order model solutions were subjected to a generalized Schmid-Leiman (S-L) transformation (Schmid & Leiman, 1957). Humphreys (1962) and others have argued that higher order model solutions should be S-L transformed to allow for clearer interpretations of the factor solution. S-L solutions also allow for more direct comparisons with corresponding bifactor solutions, as S-L coefficients represent the unique effects between indicators and latent variables. The S-L solutions were obtained via a simple multiplication procedure based on the path-tracing rule (Mulaik & Quartetti, 1997). Specifically, each subtest's first-order factor loading was multiplied by its corresponding first-order factor's second-order factor loading, which yielded the indirect *g* factor loadings. Additionally, each subtest's first-order factor loading was also multiplied by its corresponding first-order factor variable's residual variance standardized regression weight, which yielded the indirect group-level loadings (see Gignac, 2007a, for a detailed demonstration of how to obtain an S-L solution within the context of a CFA).

As the MLE chi-square associated with an SEM model has been argued to be excessively influenced by sample size (Jöreskog, 1993), the level of model fit associated with each model was evaluated based on a series of close-fit indices (aka. descriptive goodness-of-fit statistics). However, the chi-square values associated with all models (including null models) are reported for thoroughness. In choosing a series of close-fit indices, greater emphasis was placed on selecting close-fit indices that are known to include relatively greater penalties for model complexity, as an evaluation of the competing bifactor model was a goal of the investigation. Consequently, the commonly reported Standardized Root Mean Square Residual (Bentler, 1995) and Comparative Fit Index (Bentler, 1990), indices were not considered in this investigation, as the former does not include any penalty for model complexity and the latter only a weak one (Marsh, Hau, & Grayson, 2005).

Instead, the fit indices reported in this investigation included the root mean square error of approximation (RMSEA; Steiger & Lind, 1980), the Tucker-Lewis Index (TLI; Tucker & Lewis, 1973), the AIC (Akaike, 1973), and the Bayesian Information Criterion (BIC; Schwarz, 1978). RMSEA is an absolute close-fit indicator of model fit, with lower values indicative of superior model fit. Based on MacCallum, Browne, and Sugawara (1996), RSMEA values of .01, .05, and .08 were considered indicative of excellent, good, and mediocre fit, respectively. In contrast to the RMSEA, the TLI is considered an incremental

[1]Ward et al. (2011) allowed the Arithmetic subtest to load onto a third group-level factor, Crystallized Intelligence (Gc). However, given the results of Gignac (2005, 2006a), it was considered unlikely that Arithmetic would exhibit a loading onto a nested Gc factor as modeled within a bifactor modeling strategy.

fit indicator of model fit, which implies that both the null model and implied model chi-square values are used in its formulation (Marsh, Balla, & Hau, 1996). Larger values of TLI are indicative of superior model fit, with values equal to or greater than .95 indicative of good model fit (Hu & Bentler, 1999).

There are no guidelines for interpreting individual AIC and BIC values (Schermelleh-Engel, Moosbrugger, & Müller, 2003); however, in an absolute sense, smaller AIC and BIC values are indicative of superior evidence for the plausibility of a model. Within the context of evaluating competing models, Raftery (1995) suggested ΔBIC values of 0 to -2, -2 to -6, -6 to -10, and > -10 to correspond to "weak," "positive," "strong," and "very strong" indications of superior model fit, respectively.

Finally, in addition to evaluating the plausibility of each of the four competing models, the respective model-based reliability (i.e., ω_h and ω_s) estimates associated with the FSIQ and the index scores were estimated. Although a formulation approach is typically used in this context, for the purposes of efficiency, in this investigation, model-based reliability was estimated based on the implied correlation (squared) between a "phantom" composite variable and its corresponding latent variable (Fan, 2003; Raykov, 1997). In the context of this investigation, a phantom composite variable within a latent variable model is simply a representation of an equally weighted composite score (see Gignac, 2007b, for graphical representation of a phantom variable in Amos). It does not affect the level of model fit associated with the model; nor does it impact the parameter estimates (e.g., factor loadings). When squared, the implied correlation between a phantom variable and its corresponding latent variable represents the internal consistency reliability associated with the composite scores (Fan, 2003; Raykov, 1997). As applied to a bifactor model, in the context of this investigation, the implied squared correlation between the FSIQ phantom variable and the general factor latent variable is equal to ω_h. Similarly, as applied to the bifactor model, the implied squared correlation between an index phantom variable and its corresponding group-level factor latent variable is equal to ω_s. For the purposes of comparison, the internal consistency reliabilities associated with the FSIQ and index scores were estimated via coefficient α (SPSS 19.0). As per Gignac, Bates, and Jang (2007), a difference of .06 or greater between coefficient alpha and coefficient omega was considered a practically significant difference.

RESULTS

Confirmatory Factor Analyses

As can be seen in Table 1, across all four samples, the higher order and oblique factor models tended to be associated with levels of model close-fit, which

TABLE 1
Model Fit Statistics and Close-Fit Indices Associated With the CFA Models: WAIS-IV Core and Supplemental Subtests

Model	χ^2	*df*	*RMSEA (90% CI)*	*AIC (90% CI)*	*TLI*	*BIC*
Ages: 16–19						
0: Null	3,025.64	105	.264 (.256/.272)	3,055.64 (2880.12/3239.14)	.000	3,115.5
1: Higher-O.	246.75	86	.068 (.059/.079)	314.75 (271.66/365.41)	.933	450.46
2: Oblique	238.37	84	.068 (.058/.078)	310.37 (268.08/360.24)	.934	454.07
3: Bi-WAIS	147.28	75	.049 (.037/.061)	237.28 (206.57/275.56)	.965	416.90
4: Bi-CHC	150.16	74	.051 (.039/.062)	242.16 (211.04/281.26)	.963	425.77
Ages: 20–34						
0: Null	5,331.53	105	.288 (.282/.295)	5,361.53 (5127.90/5607.14)	.000	5,427.5
1: Higher-O.	298.51	86	.064 (.056/.072)	366.51 (317.40/422.80)	.950	516.00
2: Oblique	290.02	84	.064 (.056/.072)	362.02 (314.07/417.76)	.951	520.31
3: Bi-WAIS	197.21	75	.052 (.043/.061)	287.21 (249.43/332.78)	.967	485.07
4: Bi-CHC	200.62	74	.053 (.045/.062)	292.62 (254.32/340.49)	.966	494.87
Ages: 35–54						
0: Null	3,487.69	105	.284 (.276/.292)	3,517.69 (3326.25/3713.12)	.000	3,577.6
1: Higher-O.	279.28	86	.075 (.065/.085)	347.28 (300.58/403.16)	.930	482.99
2: Oblique	271.57	84	.075 (.065/.085)	343.57 (297.68/397.44)	.931	487.27
3: Bi-WAIS	160.43	75	.053 (.042/.065)	250.43 (217.73/290.71)	.965	430.04
4: Bi-CHC	172.03	74	.058 (.046/.069)	264.03 (229.33/.305.91)	.959	447.63
Ages: 55–69						
0: Null	3,740.42	105	.295 (.287/.303)	3,770.42 (3574.92/3973.90)	.000	3,830.3
1: Higher-O.	273.93	86	.074 (.064/.084)	342.93 (296.51/396.95)	.937	477.64
2: Oblique	252.69	84	.071 (.061/.081)	324.69 (280.81/376.15)	.942	468.38
3: Bi-WAIS	194.95	75	.063 (.052/.074)	284.95 (247.44/330.05)	.954	464.57
4: Bi-CHC	207.31	74	.067 (.056/.078)	299.31 (260.20/.346.00)	.948	482.92

Note. Model 0 = null model (*df* = 105); Model 0 = null model; Model 1 = higher order WAIS-IV index model; Model 2 = oblique factor WAIS-IV index model; Model 3 = bifactor WAIS-IV index model; Model 4 = bifactor CHC model; AIC = Akaike Information Criterion; NAIC = Normed Akaike Information Criterion; RMSEA = Root Mean Square Error of Approximation; TLI = Tucker-Lewis Index; Ages 16–19 (N = 400), Ages 20–34 (N = 600), Ages 35–54 (N = 400), Ages 55–69 (N = 400).

were not quite acceptable (i.e., TLI $\leq$ .93; factor loadings and latent variable correlations available upon request). In contrast to the oblique and higher order models, the bifactor model (Model 3) was associated with good levels of model close-fit across all four samples. Furthermore, across all four samples, the bifactor model was found to be better fitting than the corresponding higher order factor model (e.g., ages 20–34: ΔAIC = -79.3, ΔBIC = -30.9) and the oblique factor model (e.g., ages 20–34: ΔAIC = -74.8, ΔBIC = -35.2). Thus, based on Raftery's (1995) guidelines, the bifactor model (Model 3) was found to be associated with very strong evidence of favorability over the competing higher order and oblique factor models.

It is noteworthy that there were some important differences between the bifactor model solutions and the S-L solutions derived from the higher order model (see Table 2). The most consequential differences were observed with respect to the WM latent variable. Specifically, the S-L solution suggested equally sized, positive, and acceptably large factor loadings (.35) across all three specified WM subtests. By contrast, in the ages 55–69 sample, the bifactor model solution suggested the implausibility of a nested WM latent variable, as none of the factor loadings were statistically significant. Furthermore, across the remaining three samples, the Arithmetic subtest tended to be associated with weak and/or nonsignificant loadings on the nested WM latent variable (e.g., ages 20–34: .08, $p = .08$).

Another noteworthy difference that emerged between the S-L transformed and bifactor model solutions was relevant to the PR latent variable. Specifically, as can be seen in Table 2, all five of the specified PR subtests tended to load approximately equally (.28 to .38) onto the PR latent variable in the S-L solutions. By contrast, in the bifactor model solutions, the Matrix Reasoning (MR) and Figure Weights (FW) subtests tended to be associated with small loadings ($\approx .15$) onto the nested PR latent variable, and correspondingly larger loadings onto the general factor, in comparison with their respective general factor loadings associated with the S-L transformed higher order model solutions.

Next, the hypothesis that the WAIS-IV is more consistent with a CHC model of intelligence was tested. As can be seen in Table 1, the bifactor CHC model (Model 4) was associated with very good levels of model close-fit based on the TLI and RMSEA indices across all four samples (e.g., ages 20–34: TLI = .965, RMSEA = .053). However, as can be seen in Table 3, based on Raftery's (1995) guidelines, the bifactor WAIS-IV index model was found to be associated with "strong" to "very strong" support to suggest that it was better fitting than the competing CHC bifactor model (e.g., ages 20–34: $\Delta BIC = -9.8$). Thus, overall, the bifactor WAIS-IV index model may be suggested to be better fitting than the bifactor CHC model.

Model-Based Internal Consistency Reliability

Table 4 lists the coefficient α, ω_h, and ω_s estimates across all scales and all four samples. With respect to the FSIQ scores, the ω_h estimates across all four samples ranged between .84 and .88. By comparison, the corresponding coefficient α estimates ranged between .91 and .93. On average, the difference between the ω_h and coefficient α reliability estimates amounted to .07, which may be considered practically significant (Gignac, Bates, et al., 2007).

TABLE 2
Completely Standardized Maximum Likelihood Estimation (MLE) Solutions Associated With the Higher Order Model (Schmid-Leiman [S-L] Transformed) and the Bifactor Model: Ages 16–19, 20–34, 35–54, and 55–69

	Ages: 16–19										*Ages: 20–34*									
	S-L					*Bifactor*					*S-L*					*Bifactor*				
	g	*VC*	*PR*	*WM*	*PS*	*g*	*VC*	*PR*	*WM*	*PS*	*g*	*VC*	*PR*	*WM*	*PS*	*g*	*VC*	*PR*	*WM*	*PS*
SI	.62	.47				.60	.53				.67	.48				.69	.46			
VC	.68	.51				.68	.52				.73	.52				.72	.55			
IN	.61	.46				.67	.36				.65	.46				.66	.45			
CO	.66	.49				.65	.51				.69	.49				.72	.46			
BD	.73		.27			.68		.42			.68		.33			.66		.45		
MR	.64		.23			.66		.16			.64		.31			.69		.13		
VP	.71		.26			.67		.49			.69		.33			.66		.51		
FW	.70		.25			.74		**.10**			.69		.33			.76		.15		
PCm	.56		.20			.56		.19			.50		.24			.48		.32		
DS	.64			.48		.61			.70		.75			.29		.72			.44	
AR	.62			.46		.74			.21		.73			.28		.79			**.08**	
LN	.59			.44		.57			.46		.75			.29		.73			.43	
SS	.50				.56	.44				.68	.54				.60	.52				.67
CD	.49				.55	.47				.53	.52				.58	.53				.55
CA	.40				.45	.50				.29	.40				.44	.40				.43

(continued)

TABLE 2
(*Continued*)

	Ages: 35–54										*Ages: 55–69*									
	S-L					*Bifactor*					*S-L*					*Bifactor*				
	g	*VC*	*PR*	*WM*	*PS*	*g*	*VC*	*PR*	*WM*	*PS*	*g*	*VC*	*PR*	*WM*	*PS*	*g*	*VC*	*PR*	*WM*	*PS*
SI	.70	.49				.73	.41				.72	.43				.75	.36			
VC	.74	.52				.71	.58				.77	.45				.76	.52			
IN	.65	.46				.67	.43				.69	.41				.69	.42			
CO	.68	.48				.67	.48				.71	.42				.73	.37			
BD	.68		.38			.63		.58			.70		.32			.66		.47		
MR	.66		.37			.73		.20			.67		.30			.71		.17		
VP	.67		.38			.62		.50			.67		.30			.62		.51		
FW	.67		.38			.75		.19			.72		.32			.75		.21		
PCm	.50		.28			.46		.37			.62		.28			.62		.29		
DS	.71			.42		.67			.61		.73			.20		.71			**−.07**	
AR	.68			.40		.78			.17		.78			.21		.81			**.02**	
LN	.65			.38		.59			.49		.71			.19		.70			**−.22**	
SS	.51				.60	.47				.71	.60				.51	.56				.61
CD	.53				.62	.54				.56	.63				.54	.63				.48
CA	.31				.36	.29				.39	.44				.38	.43				.39

Note. g = general factor; VC = Verbal Comprehension; PR = Perceptual Reasoning; WM = Working Memory; PS = Processing Speed; SI = Similarities; VC = Vocabulary; IN = Information; CO = Comprehension; BD = Block Design; MR = Matrix Reasoning; VP = Visual Puzzles; FW = Figure Weights; PCm = Picture Completion; DS = Digit Span; AR = Arithmetic; LN = Letter-Number Sequencing; SS = Symbol Search; CD = Coding; CA = Cancellation; factor loadings in bold were not statistically significant ($p < .05$); the LN loading of −2.2 (ages 55–69) was associated with a Heywood case.

TABLE 3
CFA Model Fit Differences Between Competing CFA Models Across All Age Groups

	Model 3 vs. Model 1			*Model 3 vs. Model 2*			*Model 3 vs. Model 4*		
Age	*TLI*	*AIC*	*BIC*	*TLI*	*AIC*	*BIC*	*TLI*	*AIC*	*BIC*
16–19	.032	−77.5	−33.6	.034	−73.09	−34.2	.002	−4.9	−8.9
20–34	.017	−79.3	−30.9	.016	−74.8	−35.2	.001	−5.4	−9.8
35–54	.035	−96.9	−52.9	.034	−93.14	−57.2	.006	−13.6	−17.6
55–69	.017	−58.0	−13.1	.012	−39.74	−3.81	.006	−14.4	−18.4

Note. TLI = Tucker-Lewis Index; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion. Values in the table represent the results of subtracting the model fit index value of a competing model from the WAIS-IV index bifactor model fit index value; positive values associated with TLI suggest superior mode fit associated with the WAIS-IV index bifactor model; negative values associated with the AIC and BIC values suggest superior model fit associated with the WAIS-IV index bifactor model; Model 1 = WAIS-IV index higher order model; Model 2 = WAIS-IV index oblique factor model; Model 3 = WAIS-IV bifactor model; Model 4 = CHC bifactor model.

With respect to the WAIS-IV index scores, the ω_s estimates were all very low. As can be seen in Table 4, all four index scores were found to be associated with ω_s estimates less than .50, which can be contrasted to the coefficient a estimates, which ranged between .72 and .91. On average, the difference in reliability estimates amounted to .57, which greatly exceeded the practical significance criterion of .06. The PS index scores were associated with the highest level of

TABLE 4
Internal Consistency Reliabilities of WAIS-IV Index Scores as Estimated via Coefficient Alpha (α), OmegaH (ω_h), and OmegaS (ω_s): Core and Supplemental Subtests

	α					ω_h	ω_s			
Age	*FSIQ*	*VC*	*PR*	*WM*	*PS*	*FSIQ*	*VC*	*PR*	*WM*	*PS*
16–19	.91	.88	.83	.82	.72	.84	.31	.12	.28	.39
20–34	.93	.91	.84	.84	.77	.86	.29	.16	.13	.44
35–54	.92	.90	.85	.83	.74	.84	.29	.22	.24	.47
55–69	.93	.91	.86	.81	.77	.88	.22	.17	.00	.36

Note. α = coefficient α as estimated via SPSS; ω_h and ω_s = omega as estimated via the implied correlation between the WAIS-IV full scale or index latent variable and its corresponding phantom variable within Amos 19.0. FSIQ = full scale intelligence quotient; VC = Verbal Comprehension; PR = Perceptual Reasoning; WM = Working Memory; PS = Processing Speed.

ω_s estimates across all age groups (.36 to .47). By contrast, the PR index scores tended to be associated with ω_s estimates in the area of .12 to .22.[2]

DISCUSSION

The results of this investigation yielded consistent evidence in favor of a bifactor model interpretation of the WAIS-IV in comparison with the more conventional oblique and higher order models. Additionally, there was more support for considering intersubtest covariance as more consistent with the WAIS-IV index model than the CHC model. Based on the bifactor model, the Arithmetic subtest was not found to be a meaningful contributor of variance to any of the four WAIS-IV indices. Finally, the unique model-based reliabilities associated with the FSIQ scores as estimated via ω_h were found to be relatively high. By contrast, the ω_s estimates associated with the index scores were found to be very low.

The bifactor WAIS-IV model was demonstrated to be practically better fitting than the competing higher order factor model endorsed by Wechsler (2008b). The effect was consistent across all age groups and all types of close-fit indices. These results accord well with other published evaluations of the bifactor model tested on the WAIS-R and the WAIS-III (Gignac, 2005, 2006a; Golay & Lecerf, 2011; Molenaar et al., 2011). In contrast to the CFA models endorsed by Wechsler (2008b), Benson et al. (2010), and Ward et al. (2011), the very good close-fitting bifactor model endorsed in this investigation did not include any interindex cross loadings, correlated residuals, or model changes based on modification indices.

The CFA results failed to support interpreting the WAIS-IV intersubtest covariation as more aligned with the CHC model of intelligence endorsed by Ward et al., (2011). Based on Raftery's (1995) criteria for the BIC index, the bifactor WAIS model was associated with "strong" to "very strong" evidence as a better fitting model than the bifactor CHC model. Thus, the bifactor WAIS-IV index factor model may be considered better fitting than the competing CHC bifactor model.

[2]As the implied squared correlation between a phantom variable and its corresponding latent variable has not yet been applied in the bifactor context for the purposes of estimating ω_h or ω_s, we demonstrate empirically its equivalence to the formulation approaches (Formulae (2) and (3)) based on the bifactor standardized solution reported in Table 2 (ages 20–34):

$$\omega_h = \frac{94.67}{94.67 + 3.69 + 2.43 + .90 + 2.72 + 5.66} = .86.$$

As this is simply a demonstration, only the solution associated with the VC index ω_s is presented:

$$\omega_{s_{VC}} = \frac{3.69}{7.78 + 3.69 + 1.12} = .29.$$

In addition to yielding superior model fit, the bifactor model uncovered some clinically consequential factor structure results relevant to the Arithmetic subtest. Based on the Wechsler (2008b) guidelines, clinicians are instructed to use the Digit Span and Arithmetic subtests to form the WM index. However, the results of the bifactor modeling reported in this investigation suggest clearly that, if only two subtests are to be chosen as indicators of Working Memory, they should be Digit Span and Letter-Number Sequencing. That is, the Arithmetic subtest evidenced only very small loadings on the WM index factor across all age groups. For example, in the 20–34 portion of the normative sample, the Letter-Number Sequencing, Digit Span, and Arithmetic subtests were associated with loadings equal to .43, .44, and .08, respectively. Clearly, Arithmetic is only a very weak indicator of WM. Interestingly, the corresponding S-L solution failed to identify Arithmetic as a weak indicator of WM. In fact, the S-L solution suggested that Letter-Number Sequencing, Digit Span, and Arithmetic were about equally strong indicators of WM with factor loadings of .29, .29, and .28, respectively (ages 20–34). Based on a bifactor model of the WAIS-III, Gignac (2006a) also found that the Arithmetic subtest was only a very modest indicator of Working Memory (.16). As per this investigation, Gignac's (2006a) results relevant to Arithmetic were not uncovered by the corresponding and less well-fitting higher order or oblique factor models of the WAIS-III.

The PR index also evidenced noticeable differences in loadings between the S-L transformed and the bifactor model solutions. Specifically, the Matrix Reasoning and Figure Weights subtests were associated with substantially smaller loadings on the nested PR index latent variable in the bifactor solution than in the S-L transformation of the higher order model solution. Correspondingly, the Matrix Reasoning and Figure Weights subtests had larger factor loadings on the *g* factor in the bifactor model than the S-L solution. These results may be considered congruent theoretically as MR is commonly regarded as the subtest within the WAIS-IV that it is the most closely aligned with fluid intelligence and a good indicator of general intelligence (Tulsky, Saklofske, & Zhu, 2003). FW also seems to share these characteristics. In comparison with the bifactor model, it would appear that the higher order model underestimated the level of *g* saturation associated with the MR and FW subtests. This may be due to the fact that the higher order model constrains subtest general factor variance to the degree that a particular subtest is associated with the first-order latent variable it is specified to load upon. From this perspective, the higher order model may be considered a test of mediation as the first-order factors mediate the association between the subtests and the second-order general factor (Yung, Thissen, & McLeod, 1999). By contrast, the bifactor model does not imply the same test of mediation, as each subtest is specified to load onto the general factor and the respective nested factor(s) directly. For this reason, Gignac (2008) contended that the bifactor model may be considered a less theoretically complex model than the higher order model despite the fact that the bifactor model has fewer degrees of freedom. It will be noted that CFA results that support a bifactor modeling strategy are not restricted to intelligence tests. An accumulation of research has found support for the bifactor model in the area of personality as well (e.g., Chen, Hayes, Carver, Laurenceau, & Zhang, 2012; Gignac, 2007a; Gignac, 2013; Reise, 2012; Thomas, 2011).

The ω_h estimates associated with the WAIS-IV FSIQ scores were respectably high (.84 to .88) although nonetheless lower than the corresponding coefficient α estimates. The application of coefficient α to the WAIS-IV battery overestimated the internal consistency reliability associated with FSIQ by, on average, approximately .07. Brunner and Süß (2005) reported a much larger reduction in the level of internal consistency reliability associated with their total scale scores (.93 vs. .68); however, their test battery was substantially different from that of the WAIS-IV. Overall, the bifactor modeling latent variable approach to decomposing the levels of unique internal consistency reliability associated with WAIS-IV composite scores does appear to have practical effects at the total scale level. However, the level of reliability (ω_h) associated with the FSIQ scores does appear to be sufficiently high for interpretation.

In contrast to the FSIQ, the unique model-based reliabilities associated with the index scores (ω_s) were all estimated to be very low across all age groups. For example, based on the 20–34 age group, the VC, PR, WM, and PS index score ω_s values were estimated at .29, .16, .13, and .44, respectively. These results accord well with Canivez and Watkins's (2010) exploratory factor analysis of the WAIS-IV. That is, Canivez and Watkins reported that the VC, PR, WM, and PS group-level factors accounted for only 7.1%, 3.8%, 2.8%, and 5.3% of the total variance, respectively. Canivez and Watkins contended that the Wechsler (2008b) endorsed index scores should probably be considered of questionable clinical utility in applied settings, which is a contention reiterated here. With ω_s estimates substantially less than .50, the meaningful interpretation of index scores is arguably impossible. Compounding the problem of interpreting index scores as valid representations of narrow facets of cognitive ability is that the reliable variance associated with index scores is dominated by *g* factor variance. In clinical practice, it is unfeasible to decompose the reliable variance associated with index scores into their constituent parts as performed in this investigation via structural equation modeling. Thus, clinical interpretations of WAIS-IV scores should probably be restricted to the FSIQ.

It will be noted that even when using the maximum possible number of subtests (i.e., 15), the WAIS-IV indices are nonetheless defined by a maximum of only 5 subtests (PR) and as few as 3 subtests (WM and PS). Based on the principles of classical test theory, the addition of more subtests to each index should yield greater levels of internal consistency reliability, all other things remaining equal (Nunnally & Bernstein, 1994). However, based on the simulation work of Sinharay (2010), it is likely that each index would have to be defined by at least 10 subtests to add any value beyond a general factor. Thus, if the WAIS-IV is designed to measure four index scores, Sinharay's simulation research would imply that the test battery would need to be comprised of 40 subtests, which is likely much too excessive for clinical practice. Clearly, there do not appear to be any obvious or easy solutions to the problems raised in this article relevant to meaningful index score interpretations.

It should be emphasized that the effects observed in this investigation relevant to internal consistency reliability would likely be observed for other well-known cognitive ability tests. The WAIS-IV was selected simply because it is well known and has been the subject of several recent CFA investigations. Thus, future research may consider examining cognitive ability batteries such as the Stanford-Binet Intelligence Scales (Roid, 2003), the Woodcock-Johnson (Woodcock, McGrew, & Mather, 2001), and the Multidimensional Aptitude Battery (MAB; Jackson, 1998) using the same approach utilized in this investigation. It is hypothesized that all multidimensional assessments associated with a relatively strong general factor will be associated with subscales that suffer from a lack of unique internal consistency reliability as estimated via ω_s.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest, Hungary: Akadémiai Kiadó.

Bagby, R. M., Parker, J. D. A., & Taylor, G. J. (1994). The twenty-item Toronto Alexithymia Scale: I. Item selection and cross-validation of the factor structure. *Journal of Psychosomatic Research, 38,* 23–32.

Benson, N., Hulac, D. M., & Kranzler, J. H. (2010). Independent examination of the Wechsler Adult Intelligence Scale–Fourth Edition (WAIS–IV): What does the WAIS–IV measure? *Psychological Assessment, 22,* 121–130.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238–246.

Bentler, P. M. (1995). *EQS structural equations program manual*. Encino, CA: Multivariate Software Inc.

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika, 74,* 137–143.

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

Brunner, M., & Süß, H.-M. (2005). Analyzing the reliability of multidimensional measures: An example from intelligence research. *Educational and Psychological Measurement, 65,* 227–240.

Canivez, G. L., & Watkins, M. W. (2010). Investigation of the factor structure of the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS–IV): Exploratory and higher order factor analyses. *Psychological Assessment, 22,* 827–836.

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

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.

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research, 41,* 189–225.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Psychological Bulletin, 78,* 98–104.

Costello, A. B., & Osborne, J. W. (2005). Best practice and exploratory factor analysis: Four recommendations for getting the most out of your data. *Practical Assessment and Evaluation, 10,* 1–9.

Fan, X. (2003). Using commonly available software for bootstrapping in both substantive and measurement analyses. *Educational and Psychological Measurement, 63,* 24–50.

Gignac, G. E. (2005). Revisiting the factor structure of the WAIS-R: Insights through nested factor modeling. *Assessment, 12,* 320–329.

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

Gignac, G. E. (2006b). A confirmatory factor analytic examination of the factor structure of the Multidimensional Aptitude Battery (MAB): Contrasting oblique, higher-order, and nested factor models. *Educational and Psychological Measurement, 66,* 136–145.

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

Gignac, G. E. (2007b). Working memory and fluid intelligence are both identical to g?! Reanalyses and critical evaluation. *Psychology Science, 49,* 187–207.

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

Gignac, G. E. (2010). Seven-factor model of emotional intelligence as measured by Genos EI: A confirmatory factor analytic investigation based on self- and rater-report data. *European Journal of Psychological Assessment, 26,* 309–316.

Gignac, G. E. (2013). Modeling the balanced inventory of desirable responding: Evidence in favour of a revised model of socially desirable responding. *Journal of Personality Assessment*. doi: 10.1080/00223891.2013.816717

Gignac, G. E., Bates, T. C., & Jang, K. (2007). Implications relevant to CFA model misfit, reliability, and the Five Factor Model as measured by the NEO-FFI. *Personality and Individual Differences, 43,* 1051–1062.

Gignac, G. E., Palmer, B., & Stough, C. (2007). A confirmatory factor analytic investigation of the TAS-20: Corroboration of a five-factor model and suggestions for improvement. *Journal of Personality Assessment, 89,* 247–257.

Golay, P., & Lecerf, T. (2011). Orthogonal higher order structure and confirmatory factor analysis of the French Wechsler Adult Intelligence Scale (WAIS-III). *Psychological Assessment, 23,* 143–152.

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement, 66,* 930–944.

Green, S. B., & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika, 74,* 121–135.

Gustafsson, J. E., & Aberg-Bengtsson, L. (2010). Unidimensionality and the interpretability of psychological instruments. In S. E. Embretson (Ed.), *Measuring psychological constructs* (pp. 97–121). Washington, DC: American Psychological Association.

Gustafsson, J. E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research, 28,* 407–434.

Holzinger, K. J., & Swineford, R. (1937). The bifactor method. *Psychometrika, 2,* 41–54.

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.

Humphreys, L. G. (1962). The organization of human abilities. *American Psychologist, 17,* 475–483.

Jackson, D. N. (1998). *Multidimensional Aptitude Battery–II: Manual*. Port Huron, MI: Sigma Assessment Systems.

Jager, A. O., Süß, H.-H., & Beauducel, A. (1997). *Berlin Intelligence Structure–Test, Form 4*. Gottingen, Germany: Hogrefe.

Jennrich, R. I., & Bentler, P. M. (2011). Exploratory bi-factor analysis. *Psychometrika, 76,* 537–549.

Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 294–316). Newbury Park, CA: Sage.

Jöreskog, K. G., & Sörbom, D. (1989). *LISREL 7: A guide to the program and applications*. Chicago, IL: SPSS Publications.

Keith, T. Z. (2005). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 581–614). New York, NY: Guilford Press.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lucke, J. F. (2005) The α and the ω of congeneric test theory: An extension of reliability and internal consistency to heterogeneous tests. *Applied Psychological Measurement, 29,* 65–81.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1,* 130–149.

Mårdberg, B., & Carlstedt, B. (1998). The Swedish Enlistment Battery (SEB): Construct validity and latent variable estimation of cognitive abilities by CAT-SEB. *International Journal of Selection and Assessment, 6,* 107–114.

Marsh, H. W., Balla, J. R., & Hau, K. T. (1996). An evaluation of incremental fit indices: A clarification of mathematical and empirical processes. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling techniques* (pp. 315–353). Mahwah, NJ: Erlbaum.

Marsh, H. W., Hau, K. T., & Grayson, D. (2005). Goodness of fit in structural equation models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A Festschrift for Roderick P. McDonald* (pp. 225–340). Mahwah, NJ: Erlbaum.

McDonald, R. P. (1985). *Factor analysis and related methods*. Hillsdale, NJ: Erlbaum.

McGrew, K. S. (1997). Analysis of the major intelligence batteries according to a proposed comprehensive G*f*-G*c* framework. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 151–179). New York, NY: Guilford Press.

Miller, M. B. (1995). Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling. *Structural Equation Modeling, 2,* 255–273.

Molenaar, D., Dolan, C. V., & van der Maas, H. L. J. (2011). Modeling ability differentiation in the second-order factor model. *Structural Equation Modeling, 18,* 578–594.

Mulaik, S. A., & Quartetti, D. A. (1997). First order or higher order general factor? *Structural Equation Modeling, 4,* 193–211.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York, NY: McGraw-Hill.

Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research, 21,* 381–391.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25,* 111–163.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21,* 173–184.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47,* 667–696.

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2012). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment, 95,* 129–140.

Revelle, W. (2012). *Psych: Procedures for psychological, psychometric, and personality research: R package version 1.1-10.* Retrieved from http://personality-project.org/r/psych.manual.pdf

Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research, 23,* 51–67.

Roid, G. (2003). *Stanford-Binet Intelligence Scales–Fifth Edition*. Itasca, IL: Riverside.

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

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8,* 350–353.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6,* 461–464.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods for Psychological Research Online, 8,* 23–74.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74,* 107–120.

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement, 47,* 150–174.

Steiger, J. H., & Lind, J. M. (1980, May). *Statistically based tests for the number of common factors.* Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Thomas, M. L. (2011). Rewards of bridging the divide between measurement and clinical theory: Demonstration of a bifactor model for the brief symptom inventory. *Psychological Assessment, 24,* 101–113.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38,* 1–10.

Tulsky, D. S., Saklofske, D. H., & Zhu, J. (2003). Revising a standard: An evaluation of the origin and development of the WAIS–III. In D. S. Tulsky, D. H. Saklofske, G. J. Chelune, R. K. Heaton, R. J. Ivnik, R. Bornstein, … M. F. Ledbetter (Eds.), *Clinical interpretation of the WAIS–III and WMS–III* (pp. 43–91). San Diego, CA: Academic Press.

Ward, L. C., Bergman, M. A., & Hebert, K. R. (2011). WAIS-IV subtest covariance structure: Conceptual and statistical considerations. *Psychological Assessment, 24,* 328–340.

Watkins, M. W. (2010). Structure of the Wechsler Intelligence Scale for Children–Fourth Edition among a national sample of referred students. *Psychological Assessment, 22,* 782–787.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale–Revised*. New York, NY: Psychological Corporation.

Wechsler, D. (1997). *WAIS-III WMS-III technical manual*. San Antonio, TX: Psychological Corporation.

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

Wechsler, D. (2008a). *Wechsler Adult Intelligence Scale–Fourth Edition.* San Antonio, TX: Pearson Assessment.

Wechsler, D. (2008b). *Wechsler Adult Intelligence Scale–Fourth Edition: Technical and interpretive manual.* San Antonio, TX: Pearson Assessment.

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

Yung, Y.-F., Thissen, D., & McLeod, L. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika, 64,* 113–128.

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.