THE CLINICAL NEUROPSYCHOLOGIST, 2017
VOL. 31, NOS. 6–7, 1113–1126
http://dx.doi.org/10.1080/13854046.2017.1317364

Routledge
Taylor & Francis Group

The reliability of multidimensional neuropsychological measures: from alpha to omega

Marley W. Watkins

Department of Educational Psychology, Baylor University, Waco, TX, USA

ABSTRACT

Objective: To demonstrate that Coefficient omega, a model-based estimate, is more a more appropriate index of reliability than coefficient alpha for the multidimensional scales that are commonly employed by neuropsychologists. **Method:** As an illustration, a structural model of an overarching general factor and four first-order factors for the WAIS-IV based on the standardization sample of 2200 participants was identified and omega coefficients were subsequently computed for WAIS-IV composite scores. **Results:** Alpha coefficients were ≥ .90 and omega coefficients ranged from .75 to .88 for WAIS-IV factor index scores, indicating that the blend of general and group factor variance in each index score created a reliable multidimensional composite. However, the amalgam of variance from general and group factors did not allow the precision of Full Scale IQ (FSIQ) and factor index scores to be disentangled. In contrast, omega hierarchical coefficients were low for all four factor index scores (.10–.41), indicating that most of the reliable variance of each factor index score was due to the general intelligence factor. In contrast, the omega hierarchical coefficient for the FSIQ score was .84. **Conclusions:** Meaningful interpretation of WAIS-IV factor index scores as unambiguous indicators of group factors is imprecise, thereby fostering unreliable identification of neurocognitive strengths and weaknesses, whereas the WAIS-IV FSIQ score can be interpreted as a reliable measure of general intelligence. It was concluded that neuropsychologists should base their clinical decisions on reliable scores as indexed by coefficient omega.

ARTICLE HISTORY
Received 12 December 2016
Accepted 1 April 2017

KEYWORDS
Intelligence; reliability; omega; alpha; WAIS-IV

The competent practice of psychology entails adherence to professional standards, including ethical standards articulated in codes of conduct (e.g. American Psychological Association, 2002) and testing standards enumerated in the *Standards for Educational and Psychological Testing* (AERA, APA, and NCME, 2014). Among these testing Standards, 'appropriate evidence of reliability/precision' (p. 42) is vital because score validity depends on score reliability (Furr & Bacharach, 2014).

Reliability of measurement is an especially important foundation of neuropsychological practice because neuropsychologists often consider low test scores to be indicators of neuropsychological weaknesses (Decker, Hale, & Flanagan, 2013; Heyanka, Holster, & Golden, 2013). To ensure that any identified low test score is genuine and not the result of

CONTACT Marley W. Watkins Marley_Watkins@baylor.edu

measurement error, its standard error of measurement is consulted (Brooks, Strauss, Sherman, Iverson, & Slick, 2009; Crawford, Garthwaite, Longman, & Batty, 2012). Of course, the standard error of measurement for an individual examinee is statistically based on a reliability estimate for that test score (Furr & Bacharach, 2014).

Conceptually, score reliability can be considered within the classical test theory paradigm where the observed test score is hypothesized to be composed of two latent independent components: the true score plus measurement error. Error is presumed to be random but the true score is, theoretically, the mean score that would be attained if a person took the test an infinite number of times. Reliability is the ratio of true score variance to error variance (i.e. its consistency or precision).

Given that the true score is not observable, various ways to objectify it have been developed (Furr & Bacharach, 2014). Currently, the most popular quantification of score reliability is coefficient alpha (Streiner, 2003), sometimes called Cronbach's alpha (Cronbach, 1951). Alpha's popularity may be attributed to its ease of computation, reliance on a single test administration, and straightforward interpretation as percent of true score variance. However, the accuracy of coefficient alpha, like all statistical models, depends on several assumptions (Allen & Yen, 1979). Those assumptions include: (a) item errors are uncorrelated; (b) the scale measures a single construct (i.e. unidimensionality); (c) all items have the same true score variances; and (d) all items have the same relationship to the measured construct (i.e. equal factor loadings). A more technical description of parallel, tau-equivalent, and congeneric assumptions are available in measurement texts (Allen & Yen, 1979; Furr & Bacharach, 2014; Meyer, 2010).

If its basic assumptions are violated, alpha may either over or under estimate the population reliability (Cortina, 1993; Green & Hershberger, 2000; Green, Lissitz, & Mulaik, 1977; Novick & Lewis, 1967; Raykov, 2001a). Unfortunately, model assumptions are often ignored or unknown by test users (Graham, 2006; Greenland et al., 2016), including users of coefficient alpha (Henson, 2001). Further, these assumptions are unrealistic for psychological test data and likely to be violated in practice (Cho & Kim, 2015). After considering the limitations of alpha, Cronbach and Shavelson (2004, p. 403) concluded that 'I no longer regard the alpha formula as the most appropriate way to examine most data' and advocated a component of variance approach (i.e. generalizability theory).

More recently, measurement specialists have reiterated the limitations of coefficient alpha, demonstrated that its assumptions are likely violated in practice, and provided alternatives that are not dependent on such unrealistic assumptions (Green & Yang, 2009; McDonald, 1999; Raykov, 1997, 2001b; Sijtsma, 2009; Simsek & Noyan, 2013; Zinbarg, Revelle, Yovel, & Li, 2005; Zinbarg, Yovel, Revelle, & McDonald, 2006). These papers have tended to be quite technical but consistent in concluding that alpha is 'an inappropriate measure of internal consistency reliability' (Dunn, Baguley, & Brunsden, 2014, p. 402).

Model-based reliability estimates are attractive alternatives to alpha that make fewer and more realistic assumptions (Dunn et al., 2014; Reise, 2012). Critically, model-based estimates are able to properly estimate reliability for multidimensional tests where item scales and factor loadings differ (Green & Yang, 2009; Hancock & Mueller, 2001). The omega (ω) family of coefficients, first described by McDonald (1999), are the principal model-based reliability coefficients reported in current research (e.g. Canivez, Watkins, & Dombrowski, 2016). In fact, coefficient alpha is a special case of omega when alpha's assumptions are satisfied (McDonald, 1999). Especially for multidimensional constructs, omega "provides a better estimate for the

composite score [than alpha] and thus should be used (Chen, Hayes, Carver, Laurenceau, & Zhang, 2012, p. 228). Likewise, Dunn et al. (2014) advised that psychologists 'change to the routine reporting of omega in place of alpha' (p. 409) and Schweizer (2011) suggested that greater use of omega 'would be highly desirable' (p. 144).

Coefficient omega is based on a decomposition of the variance of a test within a factor analytic model into four parts: (a) a general factor with variance common to all measured variables; (b) a set of group factors (i.e. variance common to some but not all of the measured variables); (c) specific factors with variance unique to each measured variable; and (d) random error (Revelle, 2016). Specific factor variance cannot be disentangled from random error in a single test administration so they are combined (called uniqueness) in the computation of omega. Thus, omega replaces the true score theory hypothesis of true and error variance with the factor analytic conceptualization of common and unique variance.[1]

Several omega variants can be computed to describe how precisely 'total and subscale scores reflect their intended constructs' and determine 'whether subscale scores provide unique information above and beyond the total score' (Rodriguez et al., 2016a, p. 223). The most general omega coefficient is omega total (ω), 'an estimate of the proportion of variance in the unit-weighted total score attributable to all sources of common variance' (Rodriguez et al., 2016a, p. 224). High ω values indicate a highly reliable multidimensional composite. However, the amalgam of general and group variance in the computation of ω does not allow the precision of total and subscale scores to be disentangled.

ω can also be computed for each subscale score using the same computational logic. That is, the proportion of each subscale score's total variance attributed to the blend of general and group factor variance. Called omega subscale (ω_s), high values indicate a highly reliable multidimensional composite but fail to distinguish between precision of the general factor vs. precision of the group factor. Thus, omega as applied to a total score (ω) and as applied to a subscale score (ω_s) reflect the systematic variance attributable to multiple common factors. Similar to coefficient alpha, both ω and ω_s index the reliability of a multidimensional composite score.

Another omega variant, called omega hierarchical, reflects variance attributable to a common factor and is therefore a measure of the precision with which a score assesses a single construct. When applied to the general factor, ω_h is the ratio of the variance of the general factor compared to the total test variance and 'reflects the percentage of systematic variance in unit-weighted total scores that can be attributed to the individual differences on the general factor' (Rodriguez et al., 2016a, p. 224). A high ω_h coefficient indicates that the general factor is the dominant source of systematic variance in the test score. Conversely, a low ω_h coefficient indicates that group factors and/or uniqueness account for the majority of reliable variance in the test score.

When applied to group factors, the omega hierarchical variant (ω_{hs}) indicates the proportion of variance in the subscale score that is accounted for by its intended group factor (e.g. verbal comprehension factor in the VCI score, working memory factor in the WMI score, etc.) to the total variance of that subscale score and indexes the reliable variance associated with that subscale after controlling for the effects of the general factor. If ω_{hs} is low relative to ω_s, most of the reliable variance of that subscale is due to the general factor, which precludes meaningful interpretation of that subscale score as an unambiguous indicator of a group factor (Rodriguez et al., 2016b). In contrast, a robust ω_{hs} coefficient suggests that most of the reliable variance of that subscale is independent of the general factor and clinical

interpretation of an examinee's strengths and weaknesses beyond the general factor can be conducted (Brunner et al., 2012; DeMars, 2013; Reise, 2012).

The relatively recent development of omega has not yet been reflected in the technical manuals of most psychological tests (Black, Yang, Beitra, & McCaffrey, 2015), nor have omega coefficients been reported for the cognitive tests that are frequently employed by neuropsychologists (Mihura, Roy, & Graceffo, 2017). For example, neuropsychologists frequently interpret score profiles from the Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV; Wechsler, 2008a) to identify neurocognitive strengths and weaknesses (Crawford et al., 2012; Donders & Strong, 2015; Glass, Ryan, & Charter, 2010; Puente & Puente, 2013; Rabin, Paolillo, & Barr, 2016; Silver et al., 2008). Coefficient alpha reliability estimates are available for the WAIS-IV (Wechsler, 2008b) and are generally quite high (Groth-Marnat, 2009). Given these strong reliability coefficients, clinicians have been encouraged to interpret WAIS-IV score patterns, especially those at the factor index level (Groth-Marnat, 2009; Lichtenberger & Kaufman, 2009; Sattler & Ryan, 2009), and neuropsychologists routinely do so (Howieson & Lezak, 2012; Larrabee, 2014).

Establishing sufficient reliability is necessary for all educational and psychological testing applications (AERA, APA, and NCME, 2014) and especially important for evaluating the clinical utility of neuropsychological testing (Frazier, Youngstrom, Chelune, Naugle, & Lineweaver, 2004). Given that the WAIS-IV is hierarchically structured and thus multidimensional (Carroll, 1993), the statistical assumptions of coefficient alpha have likely been violated, making coefficient alpha estimates of WAIS-IV score reliability biased to an unknown extent. In turn, reliance on biased estimates of reliability may result in inaccurate clinical interpretation of WAIS-IV score patterns. Consequently, the remainder of this paper will illustrate the application of coefficient omega to the WAIS-IV to determine how precisely the WAIS-IV FSIQ and factor index scores reflect their intended constructs and whether the WAIS-IV subscale scores provide unique information above and beyond the total score.

Method

Participants

Participants were the 2,200 members of the WAIS-IV standardization sample who ranged in age from 16 to 90. The standardization sample was obtained using stratified proportional sampling across age, sex, race/ethnicity, education level, and geographic region. More detailed information is provided in the WAIS-IV Technical and Interpretive Manual (Wechsler, 2008b).

Instruments

The WAIS-IV is an individual test of intelligence that contains 10 core subtests from which a variety of composite scores are computed. First, all 10 core subtests combine to create the Full Scale IQ (FSIQ) score. Second, four factor index scores emerge from separate subtests: the Verbal Comprehension Index (VCI) and Perceptual Reasoning Index (PRI) are each composed of three subtests, whereas the Working Memory Index (WMI) and Processing Speed Index (PSI) are each composed of two subtests. Thus, a priori, the WAIS-IV is hierarchically structured and multidimensional and exhibits unequal factor loadings, violating the basic

assumptions of unidimensionality and equal factor loadings required for non-biased estimation of coefficient alpha.

Analyses

The subtest correlation matrix and standard deviations of the 10 core subtests for the total WAIS-IV standardization sample was extracted from Table 5.1 of the WAIS-V Technical and Interpretive Manual (Wechsler, 2008b) to create a covariance matrix (also published in Black et al., 2015). As omega is model-based, a higher-order confirmatory factor model consistent with that presented in Figure 5.1 of the technical manual was specified in Mplus version 7.4 (Muthén & Muthén, 2012) using maximum likelihood estimation. This model contained an overarching general factor and four first-order factors (VC, PR, WM, and PS) but honored simple structure by excluding the small (.19) complex loading of Arithmetic on the VC factor accepted by Wechsler (2008b). As expected, model fit was almost identical to that reported by Wechsler (2008b), with root mean squared error of approximation (RMSEA) of .067, comparative fit index (CFI) of .973, and Tucker-Lewis index of .961. As recommended by Carroll (1993, 1995), that hierarchical structure was then orthogonalized (Schmid & Leiman, 1957) to allow convenient computation of omega indices.

Results

The resulting WAIS-IV higher-order structure is presented in Figure 1. As expected, it was very similar to Figure 5.1 in Wechsler (2008b) and shows that the general factor exerted a strong influence on the four first-order factors that, in turn, were strongly loaded by the WAIS-IV subtests. These results are consistent with other published analyses of the

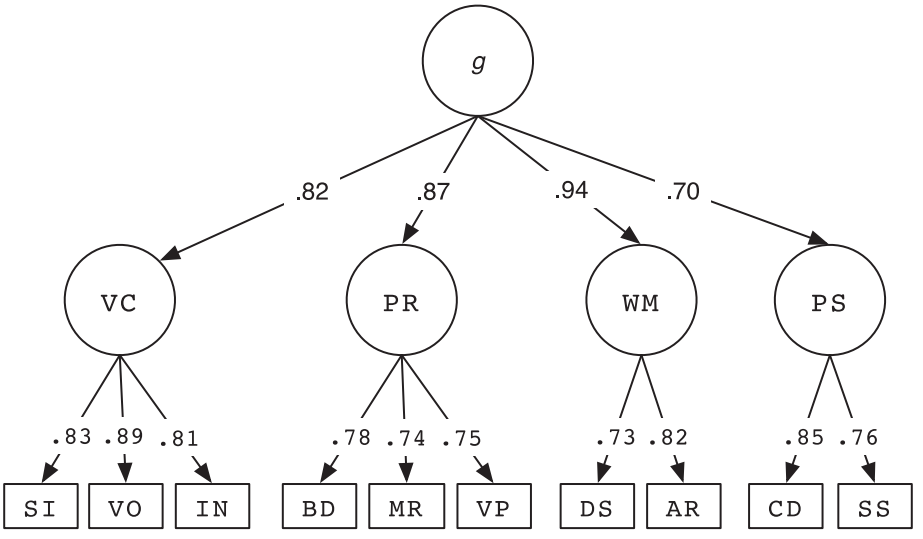

Figure 1. Higher order structure of the Wechsler Adult Intelligence Scale-Fourth Edition with its standardization sample of 2200 particpants.
Notes: SI = Similarities, VO = Vocabulary, IN = Information, BD = Block Design, MR = Matrix Reasoning, VP = Visual Puzzles, DS = Digit Span, AR = Arithmetic, CD = Coding, and SS = Symbol Search, VC = Verbal Comprehension factor, PR = Perceptual Reasoning factor, WM = Working Memory factor, PS = Processing Speed factor, *g* = General Intelligence.

standardization data (Canivez & Watkins, 2010; Gignac & Watkins, 2013; Wechsler, 2008b) and data from clinical samples (Miller, Davidson, Schindler, & Messier, 2013; Reynolds, Ingram, Seeley, & Newby, 2013).

Reliability coefficients for the WAIS-IV standardization sample were extracted from Table 4.1 of the technical manual (Wechsler, 2008b, p. 42) and are reported in Table 1.[2] A simplified omega nomenclature is applied. This terminology was adopted to reduce the confusion created by inconsistent use of ω, ω_h, ω_s, and ω_{hs} in the literature. When applied to the systematic variance attributable to multiple common factors, ω and ω_s are reported for general and group factors, respectively. In contrast, ω_h and ω_{hs} coefficients are reported as indicators of the systematic variance explained by a single general or group factor, respectively.

There is no universally accepted guideline for what constitutes adequate internal consistency reliability for clinical decisions regarding diagnosis and intervention. Various recommendations have been offered, ranging from .70 (Kline, 1998) to .96 (Kelley, 1927) with .80 to .90 most commonly recommended for decisions about individuals (Salvia, Ysseldyke, & Bolt, 2010; Thorndike & Thorndike-Christ, 2010). All WAIS-IV composite scores exhibited reliability coefficients ≥ .90 (see Table 1), suggesting that they possess adequate reliability to support clinical decisions about individuals.

However, in cases where coefficient alpha is likely biased (i.e. multidimensional measures like the WAIS-IV with unequal factor loadings), omega coefficients may be more accurate estimates than are alpha coefficients. Like alpha, there is no universally accepted guideline for acceptable or adequate levels of omega reliability for clinical decisions, but ω and ω_s coefficients should meet the same standards as alpha coefficients and ω_h and ω_{hs} coefficients should exceed .50 at a minimum but .75 would be preferred (Reise, 2012; Reise, Bonifay, & Haviland, 2013).

The degree to which composite scores, like the WAIS-IV FSIQ and index scores, are interpretable as a measure of a single common factor (i.e. FSIQ as due to general intelligence, VCI as due to verbal comprehension, PRI as due to perceptual reasoning, etc.) is indicated by the omega hierarchical coefficients in Table 1. For instance, the ω_h coefficient of .84 for the FSIQ indicates that 84% of the variance of unit-weighted FSIQ scores can be attributed to individual differences on the general intelligence factor. The square root of that ω_h (.92) is the correlation between the general factor and the observed FSIQ scores (Rodriguez et al., 2016b). A comparison of ω (variance due to general and group factors) and ω_h (variance due to general factor alone) coefficients reveals that almost all of the reliable variance in FSIQ scores can be attributed to the general factor (.84 ÷ .93 = .90). Thus, the FSIQ can confidently be interpreted as a reliable estimate of general intelligence.[3]

Table 1. Reliability estimates for Wechsler Adult Intelligence Scale-Fourth Edition composite scores.

Composite	r[a]	ω/ω_s	ω_h/ω_{hs}	H
Verbal Comprehension Index	.96	.88	.28	.47
Perceptual Reasoning Index	.95	.80	.19	.32
Working Memory Index	.94	.75	.10	.14
Processing Speed Index	.90	.79	.41	.51
Full Scale IQ	.98	.93	.84	.89

Notes: r is coefficient alpha but based on stability coefficients for the Processing Speed Index. ω and ω_s are the omega coefficients for general and group factors, respectively, and indicate the reliability of a multidimensional composite score. ω_h and ω_{hs} are the omega hierarchical coefficients for general and group factors, respectively, and reflect the reliability of the single focal factor purportedly being measured by that score. H is the construct reliability or construct replicability coefficient of Hancock and Mueller (2001).

[a]From Table 4.1 of the WAIS-IV Technical and Interpretive Manual (Wechsler, 2008b) based on the total standardization sample of 2200.

In contrast, the ω_{hs} coefficients for the four index scores ranged from .10 to .41, none meeting the minimum standard of .50 suggested by Reise (2012). The apparent reliability of index scores (i.e. α values of .90 to .96 and ω_s values of .75 to .88) was illusory because most of the explanatory power in each index score is due to the general factor. For example, the ω_s coefficient for the VCI score was .88, indicating that 88% of the variance in the VCI score was attributable to a blend of general intelligence and verbal comprehension. In contrast, the ω_{hs} coefficient of the VCI was .28, indicating that only 28% of the variance in the VCI score was attributed to the verbal comprehension construct alone. The square root of that .28 ω_{hs} coefficient (.53) is the correlation between the VC group factor and the observed VCI scores (Rodriguez et al., 2016b). A comparison of ω_s (variance due to the general and VC factors) and ω_{hs} (variance due to the VC factor alone) coefficients reveals that only a minor portion of the reliable variance in VCI scores can be attributed to the group factor (.28 ÷ .88 = .32). To interpret subscale scores with such low ω_{hs} values 'as representing the precise measurement of some latent variable that is unique or different from the general factor, clearly, is misguided' (Rodriguez et al., 2016a, p. 225).

A different perspective on WAIS-IV reliability is offered by the *H* coefficient of Hancock and Mueller (2001). Where an omega hierarchical coefficient 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 (Rodriguez et al., 2016b). Thus, *H* indicates how well a particular latent variable is represented by its indicators and is thought of as a measure of construct reliability or construct replicability (Rodriguez et al., 2016b). When *H* is low, the latent variable is not very well defined by its indicators and will tend to be unstable across studies. Table 1 reveals that only the WAIS-IV general factor was well defined, given a criterion value of .70 for *H* (Hancock & Mueller, 2001; Rodriguez et al., 2016b). Although the group factor replicability could be increased if optimally weighted composite scores were used, none reached the criterion value of .70.

Discussion

Coefficient alpha may be an inaccurate reliability index for the multidimensional scales that are commonly employed by neuropsychologists. In contrast, omega coefficients are model-based reliability estimates that make fewer and more realistic assumptions than coefficient alpha. As an illustration, omega coefficients were computed for WAIS-IV factor indices and compared to the reliability coefficients reported by Wechsler (2008b). The apparent high reliability of WAIS-IV index scores (i.e. values of .90 to .96) is illusory because most of the explanatory power in each index score is due to the general factor: The ω_{hs} coefficients for the four index scores (VCI, PRI, WMI, and PSI) indicated that each group factor (VC, PR, WM, or PS) uniquely accounted for only 28, 19, 10, and 41%, respectively, of the reliable variance of its index score. Given the imprecision with which WAIS-IV factor index scores reflected their intended constructs, their interpretation as reliable measures of an underlying group factor (i.e. verbal comprehension, perceptual reasoning, working memory, or processing speed) is misguided (Brunner et al., 2012; Canivez, 2016; Reise, 2012; Reise et al., 2013; Rodriguez et al., 2016a, 2016b). In contrast, 84% of the systematic variance in the FSIQ score was attributed to individual differences on the general factor, indicating that the FSIQ is a relatively reliable index of general intelligence not substantially affected by the multidimensionality caused by group factors (Rodriguez et al., 2016a, 2016b).

Limitations

As with all statistical indices, omega coefficients have limitations. First, their computation requires application of factor analytic models with their attendant sample size demands and interpretational complexity. This limitation is ameliorated by simulation research that found little bias in omega coefficients generated by both confirmatory and exploratory analyses as well as by principal components analyses when sample size was larger than 100 (Zinbarg et al., 2006). However, estimates of coefficient alpha are also biased by small sample sizes, with computation of both omega and alpha being more precise when sample sizes reach 300–400 (Charter, 1999). Second, omega coefficients are indices of summed unit-weighted scores and cannot be applied to scale scores that are weighted in some other way. Third, omega coefficients are estimates of internal consistency reliability and are therefore unable to detect some types of measurement error. For example, they are not sensitive to transient errors (i.e. examinees' mood or feelings on any particular day that produce consistent responses during the same assessment but inconsistent responses across different assessments). Fourth, omega coefficients are appropriate for multidimensional instruments, especially those with a hierarchical structure. These characteristics assume the source of variance lies at multiple levels (i.e. both general and group) and is orthogonal. Modern cognitive batteries, such as the WAIS-IV, with their hierarchically structured constructs are exemplars of such multidimensional instruments (Black et al., 2015; Brunner et al., 2012; Gignac & Watkins, 2013; Zinbarg et al., 2006). In contrast, instruments without a robust general factor are inappropriate candidates for estimation of reliability with omega coefficients. Fifth, there is no consensus on the optimal way to compute standard errors for omega coefficients (Kelley & Pornprasertmanit, 2016; Padilla & Divers, 2016; Zhang & Yuan, 2016). Although analytic estimates have been proposed (Raykov, 2002; Raykov & Zinbarg, 2011), their computation remains difficult and, in the case of bootstrapping methods, requires raw test data (Kelley & Cheng, 2012). However, similar ambiguity exists for the computation of standard errors for coefficient alpha (Cui & Li, 2012) so this is a shared limitation. In general, bootstrapped standard errors are probably the most accurate for both alpha and omega (Kelley & Pornprasertmanit, 2016). Regardless of method, however, lower reliability values must result in wider confidence intervals. Sixth, like all estimates of reliability, omega coefficients are based on the scores from a specific sample in a specific setting. The current study relied on scores from the WAIS-IV standardization sample. The reliability of scores from a sample of neuropsychological patients might differ. Consequently, it is important that model-based reliability be investigated among diverse samples. Finally, there is no universally accepted guideline for acceptable or adequate levels of omega reliability for clinical decisions, but it has been suggested that omega hierarchical coefficients should exceed .50 at a minimum with .75 preferable (Reise, 2012). The same uncertainty regarding coefficient alpha seems to have resulted in a clinical consensus of .80 to .90 for clinical decisions about individuals. Future research will be needed to arrive at a better consensus on guidelines for omega coefficients.

Conclusion

Notwithstanding these limitations, similar omega coefficients have been reported when different computation and analytic methods have been applied to WAIS-IV scores (Black et al.,

2015; Gignac & Watkins, 2013; Nelson, Canivez, & Watkins, 2013) and to scores from other intelligence tests (Brunner et al., 2012; Canivez & McGill, 2016; Canivez et al., 2016; Cucina & Howardson, 2016; Gomez, Vance, & Watson, 2016a, 2016b; McGeehan, Ndip, & McGill, 2017; McGill, 2016; Strickland, Watkins, & Caterino, 2015; Watkins & Beaujean, 2014). Recent simulation research revealed that high subtest score intercorrelations, as typically found in intelligence tests, always increase the reliability of the total score but reduce the distinctiveness of subscores (Bulut, Davison, & Rodriguez, 2017). Thus, the current results appear to be reasonable in the context of prior research and indicate that coefficient alpha 'misestimated reliability for the simulated and WAIS-IV examples, particularly for total scores' (Black et al., 2015, p. 469).

Measurement experts have recommended that psychologists and publishers employ coefficient omega rather than coefficient alpha because of its ability to identify the sources of test score variability and its more realistic statistical assumptions (Black et al., 2015; Chen et al., 2012; Dunn et al., 2014; Gignac, 2014; Green & Yang, 2009; Schweizer, 2011). Those recommendations were supported by the current study where omega coefficients revealed that meaningful interpretation of WAIS-IV factor index scores as unambiguous indicators of neurocognitive strengths and weaknesses may be misguided because very little reliable variance exists beyond that due to the general factor. Consequently, neuropsychologists '(a) who know what their tests can do and (b) act accordingly' (Weiner, 1989, p. 829) will base their clinical decisions (Charter & Feldt, 2001; Youngstrom & Frazier, 2013) on reliable scores as indexed by coefficient omega.

Although not currently available in test manuals, omega can be computed from exploratory or confirmatory factor results with a standalone computer program (Watkins, 2013), by hand (Brunner et al., 2012), using a so-called 'phantom variable' within confirmatory factor models (Black et al., 2015; Gignac & Watkins, 2013), and within the **R** (R Development Core Team, 2016) system. Detailed instructions for computation of omega indices, including standard errors, within the **R** system have been provided by several authors (Dunn et al., 2014; Kelley & Cheng, 2012; Revelle, 2016; Rodriguez et al., 2016b; Zhang & Yuan, 2016).

Notes

1. Formulas for omega have been presented by, among others, Brunner, Nagy, and Wilhelm (2012), Gignac (2014), McDonald (1999), Reise (2012), and Rodriguez, Reise, and Haviland (2016a, 2016b). See those publications for technical details.
2. Wechsler (2008b, p. 42) reported that 'reliability coefficients were obtained utilizing the split-half and the Cronbach's coefficient alpha methods … calculated with the formula recommended by Guilford (1954) and Nunnally and Bernstein (1994).' Gignac (2014) has suggested that inter-subtest standard alpha might be more appropriate given the multidimensional nature of the WAIS-IV. Standard alpha coefficients are more consistent with ω and ω_s coefficients in this case, but they remain dependent upon statistical assumptions, including essential tau-equivalence, whereas coefficient omega does not.
3. Omega may also be computed via bifactor confirmatory analysis and exploratory factor analysis models with orthogonalization or target bifactor rotation (Brunner et al., 2012; Reise et al., 2013; Zinbarg et al., 2005). The bifactor confirmatory method is preferred by many measurement specialists (Chen et al., 2012; Green & Yang, 2009; Reise et al., 2013; Rodriguez et al., 2016a, 2016b) but exploratory models might be useful in the absence of clear theoretical or empirical support (Zinbarg et al., 2006). In the current case, results from bifactor confirmatory analysis and exploratory factor analysis models with orthogonalization were almost identical (±.02) to those reported in Table 1. Proportionality constraints might cause some variation in results from exploratory and confirmatory models with other data (Brunner et al., 2012; Reise, 2012).

Disclosure statement

No potential conflict of interest was reported by the author.

ORCID

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

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterrey, CA: Brooks/Cole.

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

American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist, 57*, 1060–1073. doi:10.1037//0003-066X.57.12.1060

Black, R. A., Yang, Y., Beitra, D., & McCaffrey, S. (2015). Comparing fit and reliability estimates of a psychological instrument using second-order CFA, bifactor, and essentially tau-equivalent (coefficient alpha) models via AMOS 22. *Journal of Psychoeducational Assessment, 33*, 451–472. doi:10.1177/0734282914553551

Brooks, B. L., Strauss, E., Sherman, E. M. S., Iverson, G. L., & Slick, D. J. (2009). Developments in neuropsychological assessment: Refining psychometric and clinical interpretive methods. *Canadian Psychology, 50*, 196–209. doi:10.1037/a0016066

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

Bulut, O., Davison, M. L., & Rodriguez, M. C. (2017). Estimating between-person and within-person subscore reliability with profile analysis. *Multivariate Behavioral Research, 52*, 86–104. doi:10.1080/00273171.2016.1253452

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: Hogrefe.

Canivez, G. L., & McGill, R. J. (2016). Factor structure of the differential ability scales-second edition: Exploratory and hierarchical factor analyses with the core subtests. *Psychological Assessment, 28*, 1475–1488. doi:10.1037/pas0000279

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. doi:10.1037/a0020429

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

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

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

Charter, R. A. (1999). Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. *Journal of Clinical and Experimental Neuropsychology, 21*, 559–566.

Charter, R. A., & Feldt, L. S. (2001). Meaning of reliability in terms of correct and incorrect clinical decisions: The art of decision making is still alive. *Journal of Clinical and Experimental Neuropsychology, 23*, 530–537.

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

Cho, E., & Kim, S. (2015). Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods, 18*, 207–230. doi:10.1177/1094428114555994

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98–104. doi:10.1037/0021-9010.78.1.98

Crawford, J. R., Garthwaite, P. H., Longman, R. S., & Batty, A. M. (2012). Some supplementary methods for the analysis of WAIS-IV index scores in neuropsychological assessment. *Journal of Neuropsychology, 6*, 192–211. doi:10.1111/j.1748-6653.2011.02022.x

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334. doi:10.1007/BF02310555

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*, 391–418. doi:10.1177/0013164404266386

Cucina, J. M., & Howardson, G. N. (2016, November 10). 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*. Advance online. doi:10.1037/pas0000389

Cui, Y., & Li, J. (2012). Evaluating the performance of different procedures for constructing confidence intervals for coefficient alpha: A simulation study. *British Journal of Mathematical and Statistical Psychology, 65*, 467–498. doi:10.1111/j.2044-8317.2012.02038.x

Decker, S. L., Hale, J. B., & Flanagan, D. P. (2013). Professional practice issues in the assessment of cognitive functioning for educational applications. *Psychology in the Schools, 50*, 300–313. doi:10.1002/pits.21675

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

Donders, J., & Strong, C.-A. H. (2015). Clinical utility of the Wechsler Adult Intelligence Scale-Fourth Edition after traumatic brain injury. *Assessment, 22*, 17–22. doi:10.1177/1073191114530776

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

Frazier, T. W., Youngstrom, E. A., Chelune, G. J., Naugle, R. I., & Lineweaver, T. T. (2004). Increasing the reliability of ipsative interpretations in neuropsychology: A comparison of reliable components analysis and other factor analytic methods. *Journal of the International Neuropsychological Society, 10*, 578–589. doi:10.1017/S1355617704104049

Furr, R. M., & Bacharach, V. R. (2014). *Psychometrics: An introduction* (2nd ed.). Thousand Oaks, CA: Sage.

Gignac, G. E. (2014). On the inappropriateness of using items to calculate total scale score reliability via coefficient alpha for multidimensional scales. *European Journal of Psychological Assessment, 30*, 130–139. doi:10.1027/1015-5759/a000181

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

Glass, L. A., Ryan, J. J., & Charter, R. A. (2010). Discrepancy score reliabilities in the WAIS-IV standardization sample. *Journal of Psychoeducational Assessment, 28*, 201–208. doi:10.1177/0734282909346710

Gomez, R., Vance, A., & Watson, S. D. (2016a). Structure of the Wechsler Intelligence Scale for Children-Fourth Edition in a group of children with ADHD. *Frontiers in Psychology, 7*(737), 1–11. doi:10.3389/fpsyg.2016.00737

Gomez, R., Vance, A., & Watson, S. D. (2016b). Bifactor model of WISC-IV: Applicability and measurement invariance in low and normal IQ groups. *Psychological Assessment*. Advance online publication. doi:10.1037/pas0000369

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. doi:10.1177/0013164406288165

Green, S. B., & Hershberger, S. L. (2000). Correlated errors in true score models and their effect on coefficient alpha. *Structural Equation Modeling, 7*, 251–270. doi:10.1207/S15328007SEM0702_6

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement, 37*, 827–838. doi:10.1177/001316447703700403

Green, S. B., & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika, 74*, 121–135. doi:10.1007/S11336-008-9098-4

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical test, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology, 31*, 337–350. doi:10.1007/s10654-016-0149-3

Groth-Marnat, G. (2009). *Handbook of psychological assessment* (5th ed.). Hoboken, NJ: Wiley.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York, NY: McGraw-Hill.

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.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development, 34*, 177–189.

Heyanka, D. J., Holster, J. L., & Golden, C. J. (2013). Intraindividual neuropsychological test variability in healthy individuals with high average intelligence and educational attainment. *International Journal of Neuroscience, 123*, 526–531. doi:10.3109/00207454.2013.771261

Howieson, D. B., & Lezak, M. D. (2012). The neuropsychological evaluation. In S. C. Yudofsky & R. E. Hales (Eds.), *Clinical manual of neuropsychiatry* (pp. 1–26). Washington, DC: American Psychiatric Association.

Kelley, T. L. (1927). *Interpretation of educational measurements*. Chicago, IL: World Book Company.

Kelley, K., & Cheng, Y. (2012). Estimation of and confidence interval formation for reliability coefficients of homogeneous measurement instruments. *Methodology, 8*, 39–50. doi:10.1027/1614-2241/a000036

Kelley, K., & Pornprasertmanit, S. (2016). Confidence intervals for population reliability coefficients: Evaluation of methods, recommendations, and software for composite measures. *Psychological Methods, 21*, 69–92. doi:10.1037/a0040086

Kline, P. (1998). *The new psychometrics: Science, psychology, and measurement*. London: Routledge.

Larrabee, G. J. (2014). Test validity and performance validity: Considerations in providing a framework for development of an ability-focused neuropsychological test battery. *Archives of Clinical Neuropsychology, 29*, 695–714. doi:10.1093/arclin/acu049

Lichtenberger, E. O., & Kaufman, A. S. (2009). *Essentials of WAIS-IV assessment*. Hoboken, NJ: Wiley.

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

McGeehan, B., Ndip, N., & McGill, R. J. (2017). Exploring the multidimensional structure of the WASI-II: Further insights from Schmid-Leiman higher-order and exploratory bifactor solutions. *Archives of Assessment Psychology, 7*, 7–27.

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.

Meyer, J. P. (2010). *Reliability*. New York, NY: Oxford University Press.

Mihura, J. L., Roy, M., & Graceffo, R. A. (2017). Psychological assessment training in clinical psychology doctoral programs. *Journal of Personality Assessment, 99*, 153–164. doi:10.1080/00223891.2016.1201978

Miller, D. I., Davidson, P. S. R., Schindler, D., & Messier, C. (2013). Confirmatory factor analysis of the WAIS-IV and WMS-IV in older adults. *Journal of Psychoeducational Assessment, 31*, 375–390. doi:10.1177/0734282912467961

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Nelson, J. M., Canivez, G. L., & Watkins, M. W. (2013). Structural and incremental validity of the Wechsler Adult Intelligence Scale-Fourth Edition with a clinical sample. *Psychological Assessment, 25*, 618–630. doi:10.1037/a0032086

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika, 32*, 1–13. doi:10.1007/BF02289400

Nunnally, J., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: Mc-Graw-Hill.

Padilla, M. A., & Divers, J. (2016). A comparison of composite reliability estimators: Coefficient omega confidence intervals in the current literature. *Educational and Psychological Measurement, 76*, 436–453. doi:10.1177/0013164415593776

Puente, A. E., & Puente, A. N. (2013). Assessment of neuropsychological functioning. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology: Testing and assessment in clinical and counseling psychology* (Vol. 2, pp. 133–152). Washington, DC: American Psychological Association.

R Development Core Team. (2016). *R: A language and environment for statistical computing*. Retrieved from http://www.R-project.org/

Rabin, L. A., Paolillo, E., & Barr, W. B. (2016). Stability in test-usage practices of clinical neuropsychologists in the United States and Canada over a 10-year period: A follow-up survey of INS and NAN members. *Archives of Clinical Neuropsychology, 31*, 206–230. doi:10.1093/arclin/acw007

Raykov, T. (1997). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research, 32*, 329–353. doi:10.1207/s15327906mbr3204_2

Raykov, T. (2001a). Bias of coefficient α for fixed congeneric measures with correlated errors. *Applied Psychological Measurement, 25*, 69–76. doi:10.1177/01466216010251005

Raykov, T. (2001b). Estimation of congeneric scale reliability using covariance structure analysis with nonlinear constraints. *British Journal of Statistical Psychology, 54*, 315–323. doi:10.1348/000711001159582

Raykov, T. (2002). Analytic estimation of standard error and confidence interval for scale reliability. *Multivariate Behavioral Research, 37*, 89–103.

Raykov, T., & Zinbarg, R. E. (2011). Proportion of general factor variance in a hierarchical multicomponent measuring instrument: A note on a confidence interval estimation procedure. *British Journal of Mathematical and Statistical Psychology, 64*, 193–207. doi:10.1348/000711009X479714

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

Revelle, W. (2016). *An introduction to psychometric theory with applications in R*. Retrieved from http://personality-project.org/r/book/

Reynolds, M. R., Ingram, P. B., Seeley, J. S., & Newby, K. D. (2013). Investigating the structure and invariance of the Wechsler Adult Intelligence Scales, Fourth Edition in a sample of adults with intellectual disabilities. *Research in Developmental Disabilities, 34*, 3235–3245. doi:10.1016/j.ridd.2013.06.029

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

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

Salvia, J., Ysseldyke, J. E., & Bolt, S. (2010). *Assessment in special and inclusive education* (11th ed.). Belmont, CA: Wadsworth.

Sattler, J. M., & Ryan, J. J. (2009). *Assessment with the WAIS-IV*. San Diego, CA: Jerome M. Sattler.

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

Schweizer, K. (2011). On the changing role of Cronbach's α in the evaluation of the quality of a measure. *European Journal of Psychological Assessment, 27*, 143–144. doi:10.1027/1015-5759/a000069

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107–120. doi:10.1007/S11336-008-9101-0

Silver, C. H., Ruff, R. M., Iverson, G. L., Barth, J. T., Broshek, D. K., Bush, S. S., … Reynolds, C. R. (2008). Learning disabilities: The need for neuropsychological evaluation. *Archives of Clinical Neuropsychology, 23*, 217–219. doi:10.1016/j.acn.2007.09.006

Simsek, G. G., & Noyan, F. (2013). McDonald's ωt, Cronbach's α, and generalized θ for composite reliability of common factors structures. *Communications in Statistics-Simulation and Computation, 42*, 2008–2025. doi:10.1080/03610918.2012.689062

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80*, 99–103. doi:10.1207/S15327752JPA8001_18

Strickland, T., Watkins, M. W., & Caterino, L. C. (2015). Structure of the Woodcock-Johnson III cognitive tests in a referral sample of elementary school students. *Psychological Assessment, 27*, 689–697. doi:10.1037/pas0000052

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

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

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

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

Wechsler, D. (2008b). *WAIS-IV technical and interpretive manual*. San Antonio, TX: Pearson Assessment.

Weiner, I. B. (1989). On competence and ethicality in psychodiagnostic assessment. *Journal of Personality Assessment, 53*, 827–831.

Youngstrom, E. A., & Frazier, T. W. (2013). Strategies for evidence-based assessment of children and adolescents: Measuring prediction, prescription, and process. In W. E. Craighead, D. J. Miklowitz, & L. W. Craighead (Eds.), *Psychopathology: History, diagnosis, and empirical foundations* (2nd ed., pp. 36–79). Hoboken, NJ: Wiley.

Zhang, Z., & Yuan, K.-H. (2016). Robust coefficients alpha and omega and confidence intervals with outlying observations and missing data: Methods and software. *Educational and Psychological Measurement, 76*, 387–411. doi:10.1177/0013164415594658

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

Zinbarg, R. E., Yovel, I., Revelle, W., & McDonald, R. P. (2006). Estimating generalizability to a latent variable common to all of a scale's indicators: A comparison of estimators for ω_h. *Applied Psychological Measurement, 30*, 121–144. doi:10.1177/014662160527881