Psychological Methods
2016, Vol. 21, No. 2, 137–150

1082-989X/16/$12.00 http://dx.doi.org/10.1037/met0000045

Evaluating Bifactor Models: Calculating and Interpreting Statistical Indices

Anthony Rodriguez and Steven P. Reise
University of California, Los Angeles

Mark G. Haviland
Loma Linda University

Bifactor measurement models are increasingly being applied to personality and psychopathology measures (Reise, 2012). In this work, authors generally have emphasized model fit, and their typical conclusion is that a bifactor model provides a superior fit relative to alternative subordinate models. Often unexplored, however, are important statistical indices that can substantially improve the psychometric analysis of a measure. We provide a review of the particularly valuable statistical indices one can derive from bifactor models. They include omega reliability coefficients, factor determinacy, construct reliability, explained common variance, and percentage of uncontaminated correlations. We describe how these indices can be calculated and used to inform: (a) the quality of unit-weighted total and subscale score composites, as well as factor score estimates, and (b) the specification and quality of a measurement model in structural equation modeling.

Keywords: bifactor, omega, reliability, explained common variance, measurement, factor determinacy

Many psychological measures are designed primarily to scale individuals on a single construct. Psychological traits (e.g., depression, anxiety), however, often have content diverse manifestations and, thus, corresponding measures include one or more items from heterogeneous content domains to achieve content validity. As a consequence, many, if not most, commonly used assessment scales yield item response data that are more or less consistent with both unidimensional (i.e., a strong general factor) and multidimensional (i.e., two or more conceptually narrower, correlated factors) measurement models (e.g., Reise & Haviland, 2005; Reise, Moore, & Haviland, 2010).

In recent years, several authors have argued that for measures yielding multidimensional data caused by a domain structure, a bifactor measurement model (Holzinger & Harman, 1938; Holzinger & Swineford, 1937) may provide a particularly useful structural representation and be a valuable psychometric tool (Canivez, in press; Chen, Hayes, Carver, Laurenceau, & Zhang, 2012; Chen, West, & Sousa, 2006; Gignac, Palmer, & Stough, 2007; Reise, 2012; Reise, Bonifay, & Haviland, 2013; Reise et al., 2010; Reise, Morizot, & Hays, 2007). A bifactor measurement model specifies that for a given set of item responses, correlations among items can be accounted for by: (a) a general factor representing shared variance among all the items and (b) a set of group factors where variance over and above the general factor is shared among subsets of items presumed to be highly similar in content. Commonly assumed, too, is that the general and group factors are orthogonal.

The general factor represents the broad central construct an instrument intends to measure, whereas group factors represent more conceptually specific subdomain constructs. Substantively, bifactor models primarily have been used to: (a) study the partitioning of variance when it is believed that an instrument assesses both general and group sources of variance (Simms, Grös, Watson, & O'Hara, 2008), (b) control for multidimensionality, such that the measure is "essentially unidimensional" but with nuisance dimensions (Chen et al., 2006; Raykov & Pohl, 2013), (c) judge whether multidimensional item response data have a strong enough general factor to justify a unidimensional measurement model (Reise, Morizot, & Hays, 2007; Reise, Scheines, Widaman, & Haviland, 2013), and (d) determine the adequacy of a total score and what, if anything, one might gain by scoring subscales (Reise, 2012; Reise, Bonifay et al., 2013; Reise et al., 2010).

Of late, an increasing number of bifactor modeling applications have been published in major psychopathology, personality, and assessment journals. Despite this encouraging occurrence, the primary aim in most studies simply is to find a relatively better-fitting model. In the present research, we expand on this work by providing a practical guide to computing important statistical indices one can derive from bifactor models.[1] They include omega reliability coefficients, factor determinacy, construct reliability, explained common variance, and percentage of uncontaminated correlations. We describe how these indices can be calculated and used to inform the: (a) quality of unit-weighted total and subscale score composites,[2] as well as factor score estimates and (b) spec-

This article was published Online First November 2, 2015.

Anthony Rodriguez and Steven P. Reise, Department of Psychology, University of California, Los Angeles; Mark G. Haviland, Department of Psychiatry, Loma Linda University.

Correspondence concerning this article should be addressed to Anthony Rodriguez or Steven P. Reise, Department of Psychology, UCLA, 1285 Franz Hall, Los Angeles, CA 90095-1563. E-mail: anthonyr723@ucla.edu or reise@psych.ucla.edu

[1] The material also is relevant to any measure with correlated factors or a second-order structure. Data that are consistent with these models are likely to be consistent with a bifactor structure.

[2] Similar work on evaluating the relative merits of total and subscale scores in educational measurement contexts can be found in Sinharay, Haberman, and Puhan (2007) and Sinharay, Puhan, and Haberman (2011).

ification and quality of a measurement model in structural equation modeling (SEM).

Statistical Indices Derived From Bifactor Models

Before presenting these indices, we must make three clarifications. First, our presentation here concerns bifactor models fit to specific personality and psychopathology measures, which presumably have a general factor and group factors caused by clusters of items with similar content. Our guide is not directed to interpreting hierarchical structures of omnibus inventories such as the Revised NEO Personality Inventory-3 (McCrae & Costa, 2010). Second, for simplicity, we assume throughout that a well-fitting confirmatory bifactor model has been estimated, with standardized parameters and an independent cluster structure (i.e., items load on the general factor and on only one group factor). Nevertheless, the following statistics can be computed and interpreted for exploratory bifactor models as well, where the above conditions may not hold strictly. Finally, we note that the indices vary in the degree to which they are applicable only to bifactor models. Some indices—for example, the percentage of uncontaminated correlations (described below)—are bifactor-specific. Others, such as indices of factor determinacy or coefficient omega, can be calculated for several model types.

Fundamental Definitions

Because our basic latent structural model is a bifactor model, of primary importance is how an item's variance is partitioned in this framework (Figure 1).

The first two boxes represent the item variance due to a general and the group factors, respectively; these sources "cause" items to be correlated. In factor analytic terms, this is known as percent of common variance and labeled, "communality." The third box represents an item's reliable variance that is unique to each item (not shared with any other item), and the final box is variance attributable to random error. In factor analysis, specific variance and error variance cannot be separated, typically, and thus are combined in the "uniqueness" term.

What this means is that general, group, and specific sources of variance represent systematic, repeatable, and thus reliable sources of variance, whereas internal consistency estimates of reliability either lump the specific variance in the error term of the classic true plus error score model ($X = T + E$) or in the uniqueness in the common plus uniqueness factor model ($X = C + U$). As such, internal consistency indices are said to be negatively biased and only will equal the reliability when there is no item specific variance (Bentler, 2009). Note, also, that understanding the transition from the $X = T + E$ true score model to the $X = C + U$ common factor model is critical in recognizing distinctions between coefficient alpha and "model-based' reliability indices such as coefficient omega.

Figure 1. Sources of item variance.

A second important aspect of the graphic is that the bifactor model, which contains general and group sources of common variance, clearly, is multidimensional (i.e., multiple common sources of item variance and between item covariance). Given the bifactor model, in either the $X = T + E$ or the $X = C + U$ representations, T and C have two sources of variance, respectively. In turn, the percent of observed variance in a unit-weighted composite score has two sources of systematic influence (i.e., $X = T_{GEN} + T_{GRP} + E$ or $X = C_{GEN} + C_{GRP} + U$). This, however, does not imply that the "construct is multidimensional" or that the construct is a "blend of dimensions," but rather, merely that a unit-weighted composite score has more than one systematic source of variance. This not only complicates the interpretation of the composite score, it also complicates the interpretation of internal consistency indices such as alpha. One important goal, however, of the omega indices described below is to separate out the reliable variance in a composite attributable to either general or group factors.

Example Data

For the running example, we use the standardized factor loadings from a bifactor model reported in Osman et al. (2009, pp. 207–208). The Osman et al. data are based on a sample of adolescents ($n = 287$) from two inpatient units of a Midwestern state psychiatric hospital who completed the 39-item, self-report Multidimensional Anxiety Scale for Children (MASC; March, 1998). The general construct is anxiety with four subdomains, herein referred to interchangeably as group factors. These consist of the 12 physical symptom (PS) items, nine harm avoidance (HA) items, nine social anxiety (SA) items, and nine separation anxiety/panic (SP) items. In practice, these likely would be scored and reported along with the total scale score (Ivarsson, 2006; Olason, Sighvatsson, & Smari, 2004). Participants rated MASC items (e.g., "I'm tense," "I always obey," "I follow others," "I look stupid," "I'm restless") on a 4-point ordinal scale: 0 = *never true about me*, 1 = *rarely true about me*, 2 = *sometimes true about me*, and 3 = *often true about me*. Higher total scale and subscale scores reflect greater levels of general and domain specific anxiety, respectively. Item content and assignment of items to group factors (or subscales) are presented in Table 1.

In the original study, all 39 MASC items were specified to load on the general factor and on only one corresponding group factor (see Table 2); cross-loadings on group factors were not permitted, and thus all remaining loadings were constrained to zero. Moreover, group factors were orthogonal to each other and the general factor. Model fit indices were not reported for the final bifactor model;[3] however, Osman et al. (2009) concluded that support was found for a bifactor structure comprised of a general factor and four subdomains; 53.41% of the variance was attributed to the general factor, 2.39% to PS, 5.98% to HA, 3.75% to SA, and 1.43% to SP.

[3] Traditional fit indices were not reported because item parameter estimation was not based on recovering a covariance or correlation matrix. The authors used TESTFACT (Wood et al., 2003), a full-information item factor analysis program, not an SEM program.

Using the Example Data Factor Loadings

Coefficient Alpha

Although coefficient alpha (Cronbach, 1951) is the most universally reported index of internal consistency reliability, it also is the most reviled (Bentler, 2009; Revelle & Zinbarg, 2009; Sijtsma, 2009) and misunderstood (Cortina, 1993; Dunn, Baguley, & Brunsden, 2014; John & Soto, 2007; Schmitt, 1996; Yang & Green, 2011). To fully appreciate the omega indices derived from bifactor models, we begin with the basics and limits of coefficient alpha.

When coefficient alpha is computed based on a correlation matrix, and not a covariance matrix, it is "standardized coefficient alpha." It is an estimate of the percent of reliable variance in the unit-weighted total score if the items first were standardized and then added together to form a composite score. Two familiar formulas for alpha are:

Table 1
Items From the Multidimensional Anxiety Scale for Children

Item	Item #	Subscale	Content
1	1	PS	Tense
2	6	PS	Shaky
3	8	PS	Jumpy
4	12	PS	Feel weird
5	15	PS	Restless
6	18	PS	Shaking
7	20	PS	Breathing
8	24	PS	Dizzy
9	27	PS	Chest tight
10	31	PS	Heart
11	35	PS	Gastrointestinal
12	38	PS	Sweaty
13	2	HA	Permission
14	5	HA	Always obey
15	11	HA	Follow others
16	13	HA	Do right
17	21	HA	Eyes open
18	25	HA	Check first
19	28	HA	Avoid upsets
20	32	HA	Let know
21	36	HA	Check safety
22	3	SA	Laugh at me
23	10	SA	Make fun
24	14	SA	Look stupid
25	16	SA	Others think
26	22	SA	Embarrassed
27	29	SA	Ask to play
28	33	SA	Called on
29	37	SA	Public perform
30	39	SA	Shy
31	4	SP	Separation
32	7	SP	Go to camp
33	9	SP	Near mom
34	17	SP	Night light
35	19	SP	Alone
36	23	SP	Scary movies
37	26	SP	Sleep next to
38	30	SP	Car or bus
39	34	SP	Phobic

Note. PS = Physical Symptoms; HA = Harm Avoidance; SA = Social Anxiety; and SP = Separation Anxiety/Panic.

Table 2
Standardized Factor Loadings From Osman et al. (2009)

Item	General	PS	HA	SA	SP	I-ECV
1	.38	−.13	0	0	0	.89
2	.82	.10	0	0	0	.98
3	.82	.38	0	0	0	.82
4	.77	.35	0	0	0	.83
5	.74	.28	0	0	0	.87
6	.98	.02	0	0	0	.99
7	.77	.28	0	0	0	.88
8	.79	.32	0	0	0	.86
9	.66	.39	0	0	0	.74
10	.73	.19	0	0	0	.94
11	.82	.42	0	0	0	.79
12	.98	.11	0	0	0	.99
13	.24	0	.59	0	0	.14
14	.56	0	.43	0	0	.63
15	.06	0	.78	0	0	.01
16	.51	0	.71	0	0	.34
17	.46	0	.45	0	0	.51
18	.68	0	.13	0	0	.96
19	.59	0	.43	0	0	.65
20	.71	0	.17	0	0	.94
21	.60	0	.50	0	0	.59
22	.55	0	0	.75	0	.35
23	.83	0	0	.47	0	.76
24	.97	0	0	.06	0	.99
25	.90	0	0	.31	0	.89
26	.81	0	0	.51	0	.72
27	.90	0	0	.36	0	.86
28	.60	0	0	.27	0	.83
29	.78	0	0	.28	0	.89
30	.65	0	0	.17	0	.93
31	.60	0	0	0	.24	.86
32	.84	0	0	0	.36	.84
33	.66	0	0	0	.26	.86
34	.73	0	0	0	.09	.98
35	.70	0	0	0	.10	.98
36	.76	0	0	0	.28	.88
37	.74	0	0	0	.16	.95
38	.86	0	0	0	.11	.98
39	.76	0	0	0	.41	.77

Note. PS = Physical Symptoms; HA = Harm Avoidance; SA = Social Anxiety; and SP = Separation Anxiety/Panic. I-ECV is the item explained common variance.

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum \sigma_i^2}{\sigma_c^2}\right], \text{ or, } \alpha = \frac{k\bar{r}}{1+(k-1)\bar{r}} \quad (1)$$

Where k is the number of items, $\sum\sigma_i^2$ is the sum of item variances, σ_c^2 is the variance of the total composite score, and $\bar{r}$ is the average item intercorrelation. The second formula is the Spearman-Brown prophecy. When the items are standardized, item variances all are 1s; therefore, the sum of item variances is 1*k, and the variance of the total score is the sum of the correlation matrix.[4]

When a researcher does not have access to the original item response data or the original correlation matrix (which would be a polychoric correlation matrix for the Osman et al. [2009] example

[4] By the variance sum law, which states that the variance of a composite is equal to the sum of the variances of each component, plus two times the sum of the covariances across all item pairs. When the items are standardized, the correlation equals the covariance.

given the ordinal item response data), standardized alpha for the total scores can be estimated based on the bifactor loading matrix (see Table 2) as follows (see Zinbarg, Revelle, Yovel, & Li, 2005, Equation 10).[5] For the unidimensional case, alternative formulas are provided in Miller (1995) and Raykov (1997a, 1997b).

$$\alpha = \frac{k}{k-1}\left(\frac{1^T\lambda_{gen}\lambda_{gen}^T 1 - \lambda_{gen}^T\lambda_{gen} + 1^T\lambda_{grp}\lambda_{grp}^T 1 - TR(\lambda_{grp}\lambda_{grp}^T)}{Var(total)}\right) \quad (2)$$

In Equation 2, 1 is a 39 by 1 column vector of unities, λ_{gen} is the 39 by 1 matrix of loadings on the general factor, and λ_{grp}, is the 39 by 4 matrix of standardized loadings for the group factors. The *TR* term stands for trace, the sum of the diagonal values of the resulting 39 by 39 matrix. The variance of total scores can be estimated using the sum of the loadings on each factor, squared, plus the sum of the error variances.

$$Var(total) = \left(\sum\lambda_{gen}\right)^2 + \left(\sum\lambda_{grp1}\right)^2 + \left(\sum\lambda_{grp2}\right)^2 + \left(\sum\lambda_{grp3}\right)^2 + \left(\sum\lambda_{grp4}\right)^2 + \sum(1-h^2)$$
$$\alpha = \frac{39}{38}\left(\frac{745.84 - 20.50 + 39.05 - 5.27}{798.12}\right) = .98 \quad (3)$$

Using these formulas, standardized coefficient alpha for the total scale score is .98.

This value, typically, would be considered superb; almost all of the observed score variance can be attributed to "true score" variance. To what extent is this useful information, however, given that the item response data are multidimensional, which clearly is the case here? We consider this below.

Similarly, coefficient alpha can be computed for subscale scores by using the appropriate submatrices from the bifactor solution. For the PS scale, for example, we would need only the general and group factor loadings and error variances for the first 12 items. Using this procedure, the standardized alpha values for the four subscales are: .95 for PS, .88 for HA, .96 for SA, and .93 for SP. These (internal consistency) reliability values for the subscales, too, would be considered laudatory in an applied context. Again, however, we ask what do these values actually represent when the subscale scores clearly have two common sources of variance, general and group factors?

We are by no means the first authors to express concerns regarding the interpretation and value of coefficient alpha; but again, a summary of these concerns is necessary to understand what one gains by using bifactor model-derived indices. Briefly, alpha can be an accurate estimator of the ratio of true to observed scores, but its psychometric properties and interpretability depends critically on two important properties of the data: (a) they are unidimensional (one common factor such that true score variation reflects a single common source of variance) with no correlated residuals and (b) the relation between the items and true score (or the latent variable) are essentially tau equivalent (Graham, 2006). In factor analytic terms, essential tau equivalence implies that the items all have equal slope relating the latent variable to the observed responses, but the items can vary in factor intercept (i.e., means).

With this foundation,[6] we review two major concerns with alpha of particular relevance to our guide. First, if the items vary widely in factor loadings, and thus essential tau equivalence is an unreasonable assumption, then coefficient alpha seriously can underestimate the reliability of unit-weighted composite scores (Graham, 2006; Schmitt, 1996). Second, when the data are multidimensional, as in a bifactor model, alpha is influenced by all sources of common variance, and it loses its appropriateness as an indicator of how well a total or subscale score reflects a single latent variable. In fact, research has shown that alpha can be quite high (as in the present example), even when data are multidimensional, and the dimensions are orthogonal (Cortina, 1993).

To assess the interpretability of a total or subscale score in the presence of multidimensionality, in particular, multidimensionality that fits a bifactor structure, indices are required that estimate the degree to which the percent of variance in total or subscale scores is attributable to variance associated with a single latent variable. Previous authors have suggested estimating this with bifactor modeling, specifically, computing indices such as coefficient omega, omega hierarchical, and omega hierarchical subscale (Canivez, in press; Gustafsson & Aberg-Bengtsson, 2010; McDonald, 1999; Reise, 2012; Reise et al., 2010; Revelle & Zinbarg, 2009; Zinbarg et al., 2005). Following are descriptions, equations, and computations for each using the standardized factor loading matrix from the MASC.

Coefficient Omega

Coefficient omega (ω; McDonald, 1999; Revelle & Zinbarg, 2009; Zinbarg et al., 2005) is a factor analytic model-based reliability estimate. It has its origin in Jöreskog (1971). Omega estimates the proportion of variance in the observed total score attributable to all "modeled" sources of common variance (Reise, Bonifay et al., 2013; Revelle & Zinbarg, 2009). The term *modeled* is required because if a common source of variance is not included in the factor model, omega cannot account for it.

The differences between coefficients alpha and omega are that: (a) omega always is based on the factor loadings of a specific model, whereas alpha, typically, is computed based on observed variances and covariances and (b) alpha assumes equal loadings (essential tau equivalence), whereas omega is more appropriate when loadings vary (congeneric). Dunn et al. (2014) provide further review of the differences between alpha and omega (see also, Gignac & Watkins, 2013). Like alpha, the interpretability of both alpha and omega depends critically on the data being unidimensional (see Graham, 2006; McDonald, 1999). Here we present omega for multidimensional data with bifactor structure as justified in Zinbarg et al. (2005).

For the MASC, if General Anxiety (ANX) were a general factor in a bifactor structure, and PS, HA, SA, and SP items loaded on

[5] The four parts of the numerator in Equation 2 are, in order (a) the sum of general factor loadings, squared, (b) the sum of the squared general factor loadings, (c) for each group factor, the sum of loadings, squared, which are then summed, and (d) the sum of the squared group factor loadings.

[6] The properties of alpha as an estimator of reliability are complicated. For a thoughtful and extensive review and demonstration, Zinbarg, Revelle, Yovel, and Li (2005) examine alpha under different definitions of what constitutes reliable variance, under conditions of multidimensional (bifactor) and unidimensional data, and varying loadings.

four group factors, respectively, then omega for the total score is computed as:

$$\omega = \frac{\left(\sum\lambda_{gen}\right)^2 + \left(\sum\lambda_{grp1}\right)^2 + \left(\sum\lambda_{grp2}\right)^2 + \left(\sum\lambda_{grp3}\right)^2 + \left(\sum\lambda_{grp4}\right)^2}{\left(\sum\lambda_{gen}\right)^2 + \left(\sum\lambda_{grp1}\right)^2 + \left(\sum\lambda_{grp2}\right)^2 + \left(\sum\lambda_{grp3}\right)^2 + \left(\sum\lambda_{grp4}\right)^2 + \sum(1-h^2)}$$
$$\omega = \frac{745.84 + 7.34 + 17.56 + 10.11 + 4.04}{798.12} = .98 \quad (4)$$

The numerator in Equation 4 represents all of the *common* sources of unit-weighted total score variance, and the denominator is the unit-weighted total score variance (again, assuming items first are standardized). Stated differently, the denominator is all the common sources of total score variance plus the unique variance. Observe that the numerator for omega is the same as the first and third terms in the numerator for the alpha equation provided previously.

The same logic can be applied to the MASC subscales, one subset of items at a time, by using loadings and error terms corresponding to each set of items on a group factor. Using only the parameter estimates for the first 12 items, for example, omega for the PS subscale is computed as:

$$\omega_{PS} = \frac{\left(\sum\lambda_{ANX}\right)^2 + \left(\sum\lambda_{PS}\right)^2}{\left(\sum\lambda_{ANX}\right)^2 + \left(\sum\lambda_{PS}\right)^2 + \sum(1-h^2)}$$
$$\omega_{PS} = \frac{85.75 + 7.34}{85.75 + 7.34 + 3.69} = .96 \quad (5)$$

Following similar calculations, coefficient omega for the remaining subscales (HA, SA, and SP) are .90, .97, .93, respectively.

It is important to note that like alpha, omega reliability estimates reflect all sources of common variance, and item-specific variance and random error are considered error. For the total and subscale score reliability, variance of the general factor, as well as the group factors, are combined to obtain the reliability estimate. Unlike alpha, omega has the advantage that a researcher would be well aware of the multiple sources by virtue of having to specify a factor model to compute omega. To clarify the relative role of these various sources in determining composite score variance, alternative indices have been developed, namely coefficient omega hierarchical and coefficient omega hierarchical subscale (McDonald, 1999; Reise, 2012; Zinbarg et al., 2005). These coefficients take advantage of the orthogonality of group and general factors, which, in turn, allows for a unique partitioning of the common sources of variance affecting composite scores.

Omega Hierarchical

When data are suitably represented by a bifactor structure, coefficient omega hierarchical (omegaH or ω_H) is a useful model-based reliability index. Unlike alpha and omega, which estimate the proportion of variance attributable to all sources of common variance, coefficient omegaH estimates the proportion of variance in total scores that can be attributed to a single general factor, thereby, treating variability in scores due to group factors as measurement error (McDonald, 1999; Reise, Moore, & Haviland, 2013; Zinbarg, Barlow, & Brown, 1997; Zinbarg et al., 2005; Zinbarg, Yovel, Revelle, & McDonald, 2006).

Coefficient omegaH is computed by dividing the squared sum of the factor loadings on the general factor by the (*model estimated*) variance of total scores:

$$\omega_H = \frac{\left(\sum\lambda_{gen}\right)^2}{\left(\sum\lambda_{gen}\right)^2 + \left(\sum\lambda_{grp1}\right)^2 + \left(\sum\lambda_{grp2}\right)^2 + \left(\sum\lambda_{grp3}\right)^2 + \left(\sum\lambda_{grp4}\right)^2 + \sum(1-h^2)}$$
$$\omega_H = \frac{\left(\sum\lambda_{ANX}\right)^2}{\left(\sum\lambda_{ANX}\right)^2 + \left(\sum\lambda_{PS}\right)^2 + \left(\sum\lambda_{HA}\right)^2 + \left(\sum\lambda_{SA}\right)^2 + \left(\sum\lambda_{SP}\right)^2 + \sum(1-h^2)} \quad (6)$$

For the MASC, coefficient omegaH is calculated as:

$$\omega_H = \frac{745.84}{798.12} = .93$$

This value means that 93% of the variance of unit-weighted total scores can be attributed to the individual differences on the general factor. The square root of omegaH (.96) is the correlation between the general factor and the observed total scores. A comparison of omegaH (.93) with omega (.98) is critical. For the MASC, we see that almost all of the reliable variance in total scores (.93/98 = .95) can be attributed to the general factor, assumed to reflect individual differences on the trait of anxiety. Only 5% (.98 −.93) of the reliable variance in total scores can be attributed to the multidimensionality caused by the group factors. Only 2% is estimated to be due to random error. Thus, and critically, raw total scores can be interpreted as an essentially unidimensional reflection of anxiety, despite the presence of clear multidimensionality of the data.

Omega Hierarchical Subscale

A common practice in psychological research is to report coefficient alpha (less common but better, coefficient omega) for both the total scale scores as well as for subscale scores. Both alpha and omega values, however, can mislead researchers to have greater confidence than justified in the reliability of total and subscale scores as a reflection of a single latent variable. This was illustrated by the difference in omega and omegaH estimates for the MASC's general factor. Similarly, coefficient omega values for the four subscales are misleading if interpreted as reliable variance of a group factor which, in turn, may inappropriately prompt researchers to report subscale scores (and seek to identify their individual correlates).

When a bifactor model is fit to multidimensional data, however, the logic of omegaH can be extended to subscales by computing the unique variance associated with each group factor once partitioning out variance associated with a general factor. This is achieved by computing coefficient omega hierarchical subscale (omegaHS or ω_{HS}). OmegaHS is an index reflecting the reliability of a subscale score after controlling for the variance due to the general factor (Reise, Bonifay et al., 2013). To illustrate, omegaHS for the PS subscale is computed as:

$$\omega_{HS.PS} = \frac{\left(\sum\lambda_{PS}\right)^2}{\left(\sum\lambda_{ANX}\right)^2 + \left(\sum\lambda_{PS}\right)^2 + \sum(1-h^2)}$$
$$\omega_{HS.PS} = \frac{7.34}{(85.75 + 7.34 + 3.69)} = .08 \quad (7)$$

As above, only the parameter estimates for the first 12 items would be used in this computation. Coefficient omegaHS for the remaining group factors are: HA = .43, SA = .17, and SP = .08, respectively. Notice that the only computational difference between ω_{PS} and $\omega_{HS.PS}$ is that, in the numerator, the term associated with variance on the general factor is removed, leaving only the variance associated with the group factor.

Compared to the original omega coefficients for the subscales (.96, .90, .97, .93), the computed omegaHS reliability estimates are substantially reduced once controlling for a general factor (i.e., .08, .43, .17, .08). This is not surprising given that the group factors are residualized factors (representing covariances among items after removing the general factor) and that the items tend to load higher on the general than group factors (see Table 2). Clearly, once partitioning out the variance for the general factor, very little common variance remains, and, thus, subscale reliability dwindles. The apparent reliability of subscales judged by coefficient omega mostly is attributable to individual differences on the general factor.

Factor Determinacy and Construct Reliability

Latent variable modeling techniques, such as exploratory and confirmatory (restricted) factor analysis, are especially appropriate and useful for: (a) justifying the scoring of a set of items as reflecting an assumed, underlying, causal, latent variable and (b) developing measurement models for structural equation modeling (SEM), where the ultimate objective is to control for errors in measurement to better estimate the disattenuated (correcting for unreliability) relations among latent variables.

The omega values reviewed above, particularly when compared to corresponding omega values for subscales, importantly can inform: (a) on the reliability of subscale scores, (b) where the sources of reliable variance originate (general vs. group), and (c) ultimately, whether the unit-weighted scoring of subscales is justifiable. A psychometric analysis to determine the desirability of raw unit-weighted scoring of subscales, however, is not their only use.

Researchers also may ask, for example, given the results of fitting a bifactor model, what can they learn about: (a) the value of estimating factor scores,[7] especially for group factors, and then using these scores in subsequent analyses, and, in turn, (b) the value of specifying general and group factors in a measurement model in an SEM framework. We believe that these two questions can be answered adequately with studies of factor determinacy (FD; Grice, 2001) and construct reliability (Hancock, 2001; Hancock & Mueller, 2001).

We first consider factor score determinacy, its calculation and role in judging the viability of using factor scores as proxies for individual differences on a latent variable. The issue of factor score determinacy is a technically complicated one, which originated with the discovery that even in a well-fitting model, "an infinite number of ways of scoring the individuals on the factors could be derived that would be consistent with the same factor loadings" (Grice, 2001, p. 431). To the degree that factor scores are indeterminate, two researchers may estimate factor scores on the same data using two different methods, and the resulting scores may even be negatively correlated. On the other hand, to the degree that factor scores are determinate, researchers confidently can assume that individual differences on the factor score estimates are good representations of true individual differences on the factor.

In the context of a bifactor model, the determinacy of the general factor seldom will be a concern, given that all items in a measure are assumed to load saliently on this factor. What is a potential concern, however, are the group factors, which typically have few indicators and relatively lower loadings. There are several approaches to computing the degree of factor determinacy. To demonstrate, we will use the formula suggested in Beauducel (2011; Equation 4). This formula is convenient because it relies only on the model reproduced correlation matrix and not on the original correlation matrix. In the population, the correlations between factors and factor scores are:

$$FD = diag\left(\mathbf{\Phi\Lambda}^{T}\sum^{-1}\mathbf{\Lambda\Phi}\right)^{1/2} \quad (8)$$

In the above, $\mathbf{\Phi}$ is a $m \times m$ matrix of factor intercorrelations, where m is the number of factors (5×5 for the MASC). In the case of a bifactor model, this matrix always will have 1s on the diagonal and zeros elsewhere. The $\mathbf{\Lambda}$ term is a $k \times m$ matrix of standardized factor loadings where k is the number of items (39 by 5 for the MASC). $\sum$ is a $k \times k$ matrix containing the "reproduced" or model implied, correlation matrix, found by $\sum = \mathbf{\Lambda\Phi\Lambda}' + \psi$, where ψ is a 39 by 39 matrix with unique variances on the diagonal, zeros elsewhere.

When calculated, the above index provides the correlation of factor scores with the factors. Possible values range from 0 to 1, with values closer to 1 indicating better determinacy. Gorsuch (1983, p. 260) has recommended that factors score estimates only should be considered in research if their determinacy values are greater than .90. For the general and group factors (PS, HA, SA, and SP), factor determinacy values were estimated to be: .99, .86, .92, .95, and .80, respectively. These values imply that only factor scores from the general factor, HA, and SA factors are trustworthy.

Note that some analysts also report the square of these values, which indicates the percent of variance in factor scores explained by factor score estimates. In the present case, these values are .99, .74, .85, .90, and .64, respectively. Another valuable and easy to derive index is the minimum possible correlation between two sets of competing factor scores, $2\rho^2 - 1$, which are .98, .48, .71, .80, and .28, respectively. These values tell one just how different two sets of equally valid factor scores could be (Guttman, 1955; Mulaik, 1976; Steiger & Schönemann, 1978). Gorsuch (1983, p. 260) has recommended that factor scores only should be considered in research if their minimum possible correlation values are greater than .70. To some degree, this value ensures that the correlation of F1 and F2 with X, where F1 and F2 are two competing sets of factor scores and X is a criterion, at least, are in the same direction. The general factor, HA, and SA thus would be considered acceptable.

Elements of factor determinacy are relevant to understanding factor score estimates as reflections of factor scores. When measures are used in SEM, however, estimated factor scores are not needed because the disattenuated correlation between latent vari-

[7] Note that unit-weighted summed scores are, in a sense, unrefined factor score estimates (Grice, 2001).

ables is obtained directly. What is important in SEM, is how well the latent variable is represented by a particular set of items. Although certain topics, such as the need to specify at least three items to identify an orthogonal factor in SEM, are well recognized, the topic of measurement model quality of the latent variable seldom is considered carefully. To address this gap, Hancock and Mueller (2001) and Hancock (2001) introduced the term *construct reliability* (or, more recently, *construct replicability*) in an SEM context and an index to assess it.[8]

The term construct reliability, perhaps, is perplexing at first glance because the notion of reliability (true over observed score variance or repeatability of observed scores) does not fit easily into the notion of a psychological construct, which, of course, remains constant over repeated assessments and different measures. A simpler, yet still accurate, way of summarizing Hancock and Mueller's (2001) work is that construct reliability is a statistical method of judging how well a latent variable is represented by a given set of items (i.e., the quality of its indicators), and, thus, replicable across studies and ultimately useful in an SEM measurement model.

To understand the approach, consider the following index *H:*

$$H = 1/\left[1 + \frac{1}{\sum_{i=1}^{k} \frac{\lambda_i^2}{1-\lambda_i^2}}\right] \quad (9)$$

This states that for any one factor, *H* is a function of the sum of the ratios of the items' squared loading (proportion of variance explained by the latent variable) on that factor to 1 minus the squared loading (proportion of variance unexplained by the latent variable) on that factor. As the number of items increase, or the size of the factor loadings increases, *H* approaches 1; in any case it will be no smaller than the reliability of the highest loading item (i.e., no lower than the highest loading squared). Hancock and Mueller (2001) note, "the quantity represented by *H* equals the population squared multiple correlation, P^2, from regressing the construct on its indicators, that is, the proportion of variability in the construct explainable by its own indicator variables" (p. 202). "Construct" refers to a latent variable.

In contrast to omegaH and omegaHS values, which provide the correlation between a factor and a unit-weighted composite, *H* values provide the correlation between a factor and an *optimally weighted* item composite.[9] As such, this type of reliability index is more appropriate for evaluating the feasibility of specifying a measurement model in an SEM framework using a particular set of items. When *H* is low, the latent variable is not well defined by the indicators and, thus, is expected to change across studies, whereas when *H* is high, the latent variable is well defined by its indicators, which, in turn, will have more stability across studies.

For the MASC data, $H = .99, .52, .81, .70$, and .39, for the general factor and PS, HA, SA, SP, respectively. Hancock and Mueller (2001) have justified the need to meet a standard criterion of $H = .70$, and by this standard, the general factor is represented perfectly, and the subdomains HA and SA are represented well. The group factors PS and SP are not specified reliably—either more or better items are needed. In terms of specifying a measurement model, the *H* analyses suggest that only HA and SA should be included as group factors in addition to the general factor. Of course, such a model would change the *H* value of the general factor (likely trivially). The critical recognition is that with low *H*, we cannot put much trust in the estimated path coefficients between that low *H* latent variable and other latent variables because the loadings are too few or too low to reliably specify the latent variable.

Explained Common Variance

The above indices consider the sources of common variance affecting the reliable variance of total and subscale scores, the accuracy of estimated factor scores, and the appropriateness of using a set of indicators to represent a latent variable in an SEM context. We now consider a related, but distinct issue; namely, given a bifactor structure, and a reasonably strong general factor, how should the data be represented as a measurement model in an SEM—as a unidimensional measurement model or as a much more complicated, cumbersome bifactor measurement model?

The answer to this question is not always as straightforward as one might hope; simply because the data are more statistically consistent with a bifactor structure, for example, does not necessarily require that all the variables must be specified in a measurement model or that the more complex bifactor model even is necessary. It is relatively easy to find examples where authors have found that a multidimensional model fit better, yet concluded that a unidimensional measurement model likely would be adequate (Ackerman, Donnellan, & Robins, 2012; Immekus & Imbrie, 2008; Reichenheim, Moraes, Oliveira, & Lobato, 2011; Reise, Bonifay et al., 2013; Reise et al., 2007; Yang & Jones, 2008).

Many researchers appear to assume that this "unidimensional" versus "multidimensional" specification issue easily can be resolved through statistical indices of fit—if a unidimensional model fits, use it; if not, do not use it. The field of psychometrics has long dismissed such simple conceptualizations of the dimensionality of psychological data, however, and the value of statistical tests of unidimensionality (e.g., Bentler, 2009; Ten Berge & Sočan, 2004). Bentler (2009) makes this clear, "a 1-factor model hardly ever describes real data with a reasonable large *p*" (p. 141), where *p* refers to number of items.

Contemporary psychometric research in dimensionality, especially in the field of item response theory (IRT), by and large is based on evaluations of the degree of unidimensionality or multidimensionality (Reise, Moore et al., 2013). The popular DETECT statistic (Zhang & Stout, 1999), for example, has been suggested as a tool for deciding whether item response data are "unidimensional enough" or "essentially unidimensional" for the application of unidimensional IRT models. Here, we suggest an alternative index, named explained common variance (ECV; Sijtsma, 2009; Ten Berge & Sočan, 2004), which also can be used to judge the

[8] Factor determinacy and construct reliability have quite different intellectual histories. They do, however, represent two approaches to the same psychometric issue. Moreover, when the data are unidimensional, factor determinacy squared and construct reliability are equivalent. In the case of a bifactor model, different results may occur. We do not advance arguments for one over the other, however.

[9] And if the *H* value is low, then using unit-weighted item scores (i.e., not optimally weighted scores) to reflect the underlying latent variable can only be worse.

essential unidimensionality of the common variance in an item set. We are not interested in the fitting of IRT models here, but rather in deciding whether to treat the multidimensional data with a bifactor structure as essentially unidimensional in an SEM measurement model.

Prior to presenting the details of ECV, we caution that it is easy to confuse omegaH as providing an answer to the dimensionality issue, or as an indicator of general factor strength, but it does not provide either. OmegaH informs on the percent of variance in a unit-weighted composite that can be attributable to a general factor. OmegaH increases as the number of items increases (assuming they are all related to the general factor). If OmegaH is high, we can regard unit-weighted total scores to be "essentially unidimensional" in the sense that their reliable variance is influenced primarily by a single source. That is not the "dimensionality" of the data, however. We need to know the dimensionality of the data, in particular, the relative strength of the general factor, because it is the critical variable in determining whether item parameters, in both IRT and SEM, can be estimated with minimum bias.

If data are bifactor, and thus multidimensional by definition, then researchers need to consider the relative strength of those factors. In this regard, the standardized loading matrix from a bifactor model provides a very simple and elegant way to assess relative dimensional strength. A straightforward and cleaner measure of degree of essential unidimensionality is the ECV (Reise, Scheines et al., 2013; Reise et al., 2010; Sijtsma, 2008; Ten Berge & Sočan, 2004). ECV indexes variance specific to a general factor by taking the ratio of variance explained by a general factor and dividing it by the variance explained by a general and group factors where factors are assumed to be uncorrelated. The explained common variance easily is computed as:

$$ECV = \frac{\left(\sum \lambda^2_{GEN}\right)}{\left(\sum \lambda^2_{GEN}\right) + \left(\sum \lambda^2_{Grp1}\right) + \left(\sum \lambda^2_{Grp2}\right) + \left(\sum \lambda^2_{Grp3}\right) + \left(\sum \lambda^2_{Grp4}\right)} \quad (10)$$

For the MASC, ECV is:

$$ECV = \frac{20.50}{20.50 + .93 + 2.33 + 1.45 + .56} = .80$$

The computed explained common variance is .80, meaning that the general factor explains 80% of the common variance extracted with 20% of the common variance spread across groups factors.

As Reise (2012) noted, higher ECV values indicate a strong general factor, which may guide in the decision to fit a unidimensional model even to data that are multidimensional. Stated differently, when ECV is high, the factor loadings estimated in a unidimensional model may approximate well (i.e., be unbiased) the factor loadings on the general factor if a bifactor model were fit. The use of ECV in practice, however, is not simple because its relation with the parameter bias that results from model misspecification, is moderated by other factors, one of which, percentage of uncontaminated correlations, is detailed in the section following.

Finally, ECV can be computed at the item level to identify the percent of item common variance attributable to a general dimension (called I-ECV; Stucky & Edelen, 2014; Stucky, Thissen, & Edelen, 2013). These authors suggest that I-ECV can be used to select items to create a more unidimensional measure. Specifically, they recommend selecting items with I-ECV values above .80 or .85. Using the .85 criterion, 56.41% of MASC items had I-ECV values greater than or equal to .85, with values as high as .99 (items 18 "shaking," 38 "sweating," and 14 "look stupid"), 10.26% in the range of .80–.84, and the remaining values as low as .14 (item 2 "permission") and .01 (item 11 "follow others"). Almost all the PS and SP items have relatively high I-ECV values, again suggesting that these items are more pure anxiety markers and contribute little to the measurement of their respective group factors (see Table 2). On the other hand, HA and SA items tend to have relatively lower I-ECV values. These findings are consistent with the previous analyses and suggest that group factors HA and SA may have some real meaning, whereas PS and SP have little. Finally, and most interesting, some items such as item 36 ("check safety") are strong markers of both the general (.60) and group factors (.50), and with corresponding I-ECV values around .50.

Parameter Bias and Percent Uncontaminated Correlations

When researchers are concerned whether their bifactor data are "unidimensional enough" for a unidimensional IRT model or when specifying a unidimensional model in SEM, the ECV index is a valuable tool. Research has shown, however, that ECV needs to be considered within the context of the overall data structure. More specifically, if researchers are concerned about the possible biasing effects of forcing multidimensional data into a unidimensional structure, PUC, used in conjunction with ECV, can provide important information. Reise, Scheines et al. (2013) and Bonifay et al. (2015) demonstrated that parameter bias is directly related to ECV, which, in turn, is moderated by the percent of uncontaminated correlations (PUC).

To understand this, one must recognize that forcing multidimensional data, MASC data, for example, into a unidimensional measurement model is a form of model misspecification that can result in biased parameter estimates, such as factor loadings that are too high. Bias in factor loading estimates, in turn, can result in structural parameter bias. The PUC, used in conjunction with the ECV, can provide information on the conditions under which this bias is more or less acute.

To understand PUC, and the role it plays in influencing parameter bias, consider the MASC data where there are $(39 \times 38)/2 = 741$ unique correlations. Within each group factor, item correlations are contaminated by variance attributed to both the general and group factor resulting in $[(12 \times 11)/2 + 3(9 \times 8)/2)] = 174$ contaminated (by multidimensionality) correlations. The correlations between items from different group factors reflect general factor variance only, and there are 741–174 = 567 of those uncontaminated (by multidimensionality) correlations. Thus, the PUC is 567/741 = .77. What that means, in an applied sense, is that the overwhelming majority of 741 correlations inform directly on the general factor, which is the target trait the instrument was designed to assess.

Reise, Scheines et al. (2013) and Bonifay, Reise, Scheines, and Meijer (2015) showed that as PUC increases, the magnitude of the ECV value becomes less and less important in determining the potential for bias when a unidimensional model is fit to multidi-

mensional data with bifactor structure. The technical reasons for this are complicated, but the simple explanation is that as PUC increases, the general trait in the bifactor model becomes more and more similar to the single trait estimated in a unidimensional model, especially when ECV is high. By definition, PUC becomes large when there are many items and many small group factors. The smaller the group factors, the more correlations there are that are influenced by only a single latent variable.

To demonstrate the use of ECV and PUC, consider the MASC, with values of .80 and .77, respectively, on these indices. Under such conditions, we expect very little difference in the factor loadings between a unidimensional model and the general factor in a bifactor model. To illustrate, we used the reproduced correlation matrix from the MASC described previously and fit both unidimensional and bifactor models. We then computed the relative parameter bias as the difference between an item's loading in the unidimensional solution and its general factor loading in the bifactor (i.e., the truer model), divided by the general factor loading in the bifactor. We found that the average relative bias across items was 2%. According to Muthén, Kaplan, and Hollis (1987), parameter bias less than 10–15% is acceptable and poses no serious concern.

As a consequence, a unidimensional measurement model in the context of SEM well may suffice for the MASC, even though such a model would not provide a good statistical fit to the data. Another, and, perhaps, far easier approach would be to form parcels by subdomain and then include only those four parcels in the SEM measurement model to represent anxiety (Little, Rhemtulla, Gibson, & Schoemann, 2013). In theory, if data are perfectly bifactor, then forming parcels based on content domains should yield a unidimensional set of indicators. Explorations of forming parcels from real data, which cannot be expected to fit perfectly, is a much needed research endeavor (Bandalos, 2002; Bandalos & Finney, 2001; Coffman & MacCallum, 2005; Little, Cunningham, Shahar, & Widaman, 2002; Sterba, 2011; Sterba & MacCallum, 2010).

Discussion

Bifactor measurement models—latent structural models that propose a single general factor as well as multiple uncorrelated group factors—are commonly being applied to personality and psychopathology measures (Reise, 2012). Most bifactor model applications, however, are limited to demonstrations of "superior fit" and partitioning sources of item response variance. In the present article, we have shown how bifactor modeling can be used to more thoroughly evaluate an instrument's psychometric properties and with a running example using solely the standardized loading matrix from the MASC (Osman et al., 2009) study,[10] how statistics derived from bifactor models can aid in practical decision making.

The indices we presented roughly can be divided into two types: those that inform on (a) properties of total and subscale scores derived from an instrument and (b) the use of a measure in an SEM framework. Following, we briefly review the definition of each index and the practical issues they each address (with comments on the MASC results).

MASC Evaluation Through a Bifactor Lens

Practical issue: Judging the reliability of unit-weighted composite scores.

Coefficient alpha. Alpha is an estimate of the reliability (true score variance over observed score variance) of unit-weighted test scores. If data are unidimensional, with no correlated residuals, alpha can be a lower bound estimate of reliability. Alpha assumes essential tau equivalence (equal factor loadings) and depends on the average item intercorrelation and the number of items (as these values increase alpha also increases). In the MASC, alpha was very high for total scores, as well as for the four subscale scores: .99 for Anxiety, .95 for PS, .88 for HA, .96 for SA, and .93 for SP. This high internal consistency is partly attributable to many content redundant items within group factors, which inflate correlations.

To what degree the correlation between content-similar items is due to a common trait or traits(s) versus an artifact of just asking the question more than once is a complicated topic, which would require a separate paper. For our present purposes, we suggest checking this by estimating a series of models, each time eliminating one or two items from sets that appear overly content redundant. If the loadings truly reflect the relation with a common trait, they should be invariant as to other items in the model. If the loadings change substantially when items are removed, this can be taken as a sign that the latent variable is overly influenced by sets of content redundant items. The logic underlying this method is well described by Kievit et al (2011, p. 71). ". . . in correctly specified reflective models, latent variables should be referentially stable. That is to say that the addition or deletion of an indicator may alter the accuracy by which the attribute is measured but not the nature of the attribute (latent variable) itself. With regard to the measurement of *g*, Spearman called this characteristic *indifference of the indicators* (Spearman, as cited in Jensen, 1998)."

Coefficient omega. This is a factor analytic model-based estimate of the reliability (true score variance over observed score variance) of unit-weighted test scores. Omega is the model-based analogue of coefficient alpha, except that it is appropriate for congeneric tests (varying factor loadings). Its value is influenced by all modeled sources of common variance. Like any internal consistency estimate, it is a negatively biased estimate of reliability because it includes item specific variance as error (Bentler, 2009). In the MASC, omega was very high for both total and subscale scores: .98, .96, .90, .97, and .93, respectively. The same caveat about item content redundancy expressed for alpha, applies equally to interpreting omega.

Coefficient omega hierarchical (omegaH). This is the percent of total score variance attributable to a single general factor. The square root provides the correlation of raw scores with the general factor. For the MASC, omegaH was .93; when compared to omega of .98, it is clear that the overwhelming majority of reliable variance in the total scores is attributable to the general factor. We concluded that raw scores essentially are univocal indicators of the general factor and only trivially affected by multidimensionality caused by group factors.

[10] In the Appendix we provide an annotated example analysis using the R (R Development Core Team, 2014) psych library 1.4.5 (Revelle, 2014).

Coefficient omega hierarchical subscale (omegaHS). This is the percent of subscale score variance attributable to a group factor, after removing the reliable variance due to the general factor. The square root provides the correlation of raw unit-weighted subscale scores with the group factor. For the MASC, omegaHS values were low for the four subscales, especially when compared to their corresponding omega values. The majority of reliable variance in subscale scores was attributable to the general factor, which precludes meaningful interpretation of MASC subscale scores as unambiguous indicators of a group factor. For the MASC, the only way to increase omegaH would be to find items that were "pure" indicators of the group factor and had low relations with the general factor—a task, in fact, that may not be possible.

Practical issue: Using a set of items to compute factor scores or identifying a latent variable in an SEM context.

Factor determinacy. For any factor model, there are an infinite set of equally valid factor score estimates. When factor determinacy—the correlation between factor score estimates and factors—is high, this is not a problem, because any competing set of factor scores will yield nearly identical results. When factor determinacy is low, one cannot be confident in the factor score estimates (or unit-weighted scores), because competing estimates may yield completely different results. In the MASC, the general factor and the HA and SA group factors emerged as acceptably determinate, whereas the PS and SP group factors displayed low determinacy. In general, seldom would we expect determinacy problems with general factors defined by many items, but we are more cautious in our expectations for group factors.

Construct reliability (or construct replicability). This is the quality of the measurement model defined by a particular set of items or how well the items reflect or account for the variance of the latent variable. When the data are unidimensional, FD^2 and H are equivalent, but when data are bifactor, the values may differ and lead to different conclusions. If a set of items has low construct reliability, which will occur when there are few items with relatively low loadings, the latent variable may be identified but not reliably specified. In such cases, we would not expect SEM results to replicate well. When construct reliability is high, our expectations are just the opposite.

For the MASC data, the *H* values were .99, .52, .81, .70, and .39, for the general factor and PS, HA, SA, SP, respectively. Because *H* values are interpretable as a type of "reliability" coefficient, we can conclude that HA and SA are "acceptable" using the .70 benchmark. The *H* values for PS and SP do not meet this criterion, however, and these results would lead us be highly suspect of any structural equation model that included PS or SP as group factors.

Note, however, the critical role of high loadings in determining *H* values. In the MASC, several items had loadings greater than .90 on the general factor—thus, causing high construct reliability. As noted above in the Alpha and Omega sections, if those high loadings merely reflect the effects of shared variance caused by the repeated item content, rather than each item's relation with a single common latent variable, then that construct reliability is illusory (or, the latent variable reflects nothing of substance, but the items are indicating it well).

The *H* values for all four group factors (.52, .81, .70, and .39) are much higher relative to corresponding omegaHS values (.08, .43, .17, .08), suggesting that unit-weighted subscale scores are poor, but using these items to specify a group factor may be feasible, at least in the case of the second and third group factors, respectively. These results are not at all contradictory. The values of omegaHS, are strongly affected by the absolute and relative sizes of item loadings on the general and group factors, and when items load strongly on the general factor, omegaHS values will be low. *H* values are affected only by the items' loadings on the factor of interest. Finally, we observed that for the MASC, *H* values for subdomain (group or residualized) factors are expected to be much lower than their corresponding values in, say, a four correlated factors model (not residualized). This phenomenon is completely analogous to comparing omega with omegaHS values—the latter are inevitably smaller than the former due to the residualized nature of group factors.

Practical issue: Deciding whether multidimensional (bifactor) data are "unidimensional enough" to specify a unidimensional measurement model in an SEM context.

Explained common variance (ECV). ECV is the percent of common variance explained by the general factor. This is a degree of unidimensionality index and is directly related to the relative strength of the general factor. It is useful, in conjunction with PUC, in deciding whether the data are essentially unidimensional, such that fitting a unidimensional latent variable model will not lead to seriously biased parameter estimates. The ECV for the MASC was .80, indicating a strong general factor. ECV should not be confused with omegaH (see above definition); ECV does not necessarily increase as items are added to a measure, but omegaH does, assuming items are related to the general factor.

Item explained common variance (I-ECV). For a single item, I-ECV is the percent of common variance (communality) due to the general factor. It is suggested as a useful index for scale construction and refinement (Stucky & Edelen, 2014). Items with high I-ECV are good candidates for inclusion on a measure if the goal is to create a unidimensional (one common factor) item set. For the MASC, many items had very large I-ECV values suggesting they are relatively pure markers of anxiety and not their corresponding group factor. Many of the items in the PS and SP group factors match this description. On the other hand, items from HA and SA group factors also loaded highly on the general factor, but they also tended to have substantial loadings on their corresponding group factor.

Percent uncontaminated correlations (PUC). PUC is the number of unique correlations in a correlation matrix that are influenced by a single factor divided by the total number of unique correlations. The higher the PUC, the more the matrix is saturated with information relevant to estimating the parameters of a single factor and the less likely the parameter estimates in a unidimensional model will be biased. In the MASC, PUC was .77, and when combined with a strong general factor (ECV = .80), one reasonably can conclude that the common variance is essentially unidimensional. When we compared the parameter estimates from the general factor in a bifactor model with those from a unidimensional model, the relative difference was only 2%. Whereas a 39-item unidimensional measurement model may be acceptable, it probably is more parsimonious to form four parcels based on the item content subdomains and simply use those four continuous indicators to identify the single latent variable. Without detailing the mathematics of parcels (but see Sterba, 2011), collapsing sets of items (e.g., 39) that are content

homogeneous creates a new, much smaller subset of observed variables (in this case 4), where the common variance among them should reflect only the general factor. Parcels also are more likely to be reasonably considered continuous and normally distributed. In other words, parceling can eliminate the multidimensionality from the indicators and make the data more amenable to SEM analyses using maximum likelihood methods.

Conclusion

We presented indices that one can derive from the results of a bifactor model; each provides information about various aspects of a measure's psychometric properties. We intentionally have sidestepped the theoretical debates about the virtues of the bifactor model versus competing models; moreover, we have not offered technical advice regarding estimation methods or firm benchmarks for judging model fit. Those issues are well beyond the present scope.

We believe that this set of indices offers tremendous potential to assist scale developers and evaluators, as well as those who use the scales in research and clinical practice. We thank a reviewer for helping us clarify our recommendation that these indices become routinely reported in research articles and in test technical manuals so that researchers and clinicians have information available to adequately judge whether a measure has acceptable true score variance. Further, we argue that these indices will have high generalizability, given that any set of items (or subtests such as in tests of intelligence and achievement) consistent with a second-order or correlated factors model conceivably could be modeled as bifactor. Finally, we also believe in their potential to inform construct development and understanding, even for researchers who do not entirely endorse the bifactor model as a valid representation of the structure of psychological traits.

References

Ackerman, R. A., Donnellan, M. B., & Robins, R. W. (2012). An item response theory analysis of the Narcissistic Personality Inventory. *Journal of Personality Assessment, 94,* 141–155. http://dx.doi.org/10.1080/00223891.2011.645934

Bandalos, D. L. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling, 9,* 78–102. http://dx.doi.org/10.1207/S15328007SEM0901_5

Bandalos, D. L., & Finney, S. J. (2001). Item parceling issues in structural equation modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 269–296). Mahwah, NJ: Erlbaum.

Beauducel, A. (2011). Indeterminacy of factor score estimates in slightly misspecified confirmatory models. *Journal of Modern Applied Statistical Methods, 10,* 583–598.

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika, 74,* 137–143. http://dx.doi.org/10.1007/s11336-008-9100-1

Bonifay, W. E., Reise, S. P., Scheines, R., & Meijer, R. R. (2015). When are multidimensional data unidimensional enough for structural equation modeling?: An evaluation of the DETECT multidimensionality index. *Structural Equation Modeling*. Advance online publication. http://dx.doi.org/10.1080/10705511.2014.938596

Canivez, G. L. (in press). 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*. Gottingen, Germany: Hogrefe Publishers.

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. http://dx.doi.org/10.1111/j.1467-6494.2011.00739.x

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. http://dx.doi.org/10.1207/s15327906mbr4102_5

Coffman, D. L., & MacCallum, R. C. (2005). Using parcels to convert path analysis models into latent variable models. *Multivariate Behavioral Research, 40,* 235–259. http://dx.doi.org/10.1207/s15327906mbr4002_4

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

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

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. http://dx.doi.org/10.1111/bjop.12046

Gignac, G. E., Palmer, B. R., & 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. http://dx.doi.org/10.1080/00223890701629730

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. http://dx.doi.org/10.1080/00273171.2013.804398

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

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. http://dx.doi.org/10.1177/0013164406288165

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods, 6,* 430–450. http://dx.doi.org/10.1037/1082-989X.6.4.430

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. http://dx.doi.org/10.1037/12074-005

Guttman, L. (1955). The determinacy of factor score matrices with implications for five other basic problems of common factor theory. *British Journal of Statistical Psychology, 8,* 65–81. http://dx.doi.org/10.1111/j.2044-8317.1955.tb00321.x

Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and MIMIC approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika, 66,* 373–388. http://dx.doi.org/10.1007/BF02294440

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future—A Festschrift in honor of Karl Jöreskog* (pp. 195–216). Lincolnwood, IL: Scientific Software International.

Holzinger, K. J., & Harman, H. H. (1938). Comparison of two factorial analyses. *Psychometrika, 3,* 45–60. http://dx.doi.org/10.1007/BF02287919

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 2,* 41–54. http://dx.doi.org/10.1007/BF02287965

Immekus, J. C., & Imbrie, P. K. (2008). Dimensionality assessment using full-information item bifactor analysis for graded response data: An

illustration with the State Metacognition Inventory. *Educational and Psychological Measurement, 68,* 695–709. http://dx.doi.org/10.1177/0013164407313366

Ivarsson, T. (2006). Normative data for the Multidimensional Anxiety Scale for Children (MASC) in Swedish adolescents. *Nordic Journal of Psychiatry, 60,* 107–113. http://dx.doi.org/10.1080/08039480600588067

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger/Greenwood.

John, O. P., & Soto, C. J. (2007). The importance of being valid. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 461–494). New York, NY: Guilford Press.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36,* 109–133. http://dx.doi.org/10.1007/BF02291393

Kievit, R. A., Romeijn, J. W., Waldorp, L. J., Wicherts, J. M., Scholte, H. S., & Borsboom, D. (2011). Mind the gap: A psychometric approach to the reduction problem. *Psychological Inquiry, 22,* 67–87. http://dx.doi.org/10.1080/1047840X.2011.550181

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9,* 151–173. http://dx.doi.org/10.1207/S15328007SEM0902_1

Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods, 18,* 285–300. http://dx.doi.org/10.1037/a0033266

March, J. (1998). *Manual for Multidimensional Anxiety Scale for Children (MASC)*. Toronto, ON: Multi-Health Systems Inc.

McCrae, R. R., & Costa, P. T. (2010). *NEO Inventories for the NEO Personality Inventory-3 (NEO PI-3), NEO Five-Factor Inventory-3 (NEO-FFI-3) and NEO Personality Inventory-revised (NEO PI-R): Professional manual*. Lutz, FL: Psychological Assessment Resources.

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

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. http://dx.doi.org/10.1080/10705519509540013

Mulaik, S. A. (1976). Comments on "the measurement of factorial indeterminacy". *Psychometrika, 41,* 249–262. http://dx.doi.org/10.1007/BF02291842

Muthén, B. O., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52,* 431–462. http://dx.doi.org/10.1007/BF02294365

Ólason, D. T., Sighvatsson, M. B., & Smári, J. (2004). Psychometric properties of the Multidimensional Anxiety Scale for Children (MASC) among Icelandic schoolchildren. *Scandinavian Journal of Psychology, 45,* 429–436. http://dx.doi.org/10.1111/j.1467-9450.2004.00424.x

Osman, A., Williams, J. E., Espenschade, K., Gutierrez, P. M., Bailey, J. R., & Chowdhry, O. (2009). Further evidence of the reliability and validity of the Multidimensional Anxiety Scale for Children (MASC) in psychiatric inpatient samples. *Journal of Psychopathology and Behavioral Assessment, 31,* 202–214. http://dx.doi.org/10.1007/s10862-008-9095-z

Raykov, T. (1997a). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21,* 173–184. http://dx.doi.org/10.1177/01466216970212006

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

Raykov, T., & Pohl, S. (2013). Essential unidimensionality examination for multicomponent scales: An interrelationship decomposition approach. *Educational and Psychological Measurement, 73,* 581–600. http://dx.doi.org/10.1177/0013164412470451

R Development Core Team. (2014). R: A language and environment for statistical computing [Software]. Vienna, Austria: R Foundation for Statistical Computing, Retrieved from http://www.R-project.org

Reichenheim, M. E., Moraes, C. L., Oliveira, A. S. D., & Lobato, G. (2011). Revisiting the dimension structure of the Edinburgh Postnatal Depression Scale (EPDS): Empirical evidence for a general factor. *BMC Medical Research Methodology, 11,* 93. http://dx.doi.org/10.1186/1471-2288-11-93

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47,* 667–696. http://dx.doi.org/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. http://dx.doi.org/10.1080/00223891.2012.725437

Reise, S. P., & Haviland, M. G. (2005). Item response theory and the measurement of clinical change. *Journal of Personality Assessment, 84,* 228–238. http://dx.doi.org/10.1207/s15327752jpa8403_02

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

Reise, S. P., Moore, T. M., & Haviland, M. G. (2013). Applying unidimensional item response theory models to psychological data. In K. Geisinger (Ed.), *APA handbook of testing and assessment in psychology: Vol. 1. Test theory and testing and assessment in industrial and organizational psychology* (pp. 101–119). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/14047-006

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16,* 19–31. http://dx.doi.org/10.1007/s11136-007-9183-7

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement, 73,* 5–26. http://dx.doi.org/10.1177/0013164412449831

Revelle, W. (2014). psych: Procedures for personality and psychological, psychometric, and psychological research [Software]. Evanston, IL: Northwestern University. Retrieved from http://CRAN.R-project.org/package=psychVersion=1.4.5

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika, 74,* 145–154. http://dx.doi.org/10.1007/s11336-008-9102-z

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8,* 350–353. http://dx.doi.org/10.1037/1040-3590.8.4.350

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74,* 107–120. http://dx.doi.org/10.1007/s11336-008-9101-0

Simms, L. J., Grös, D. F., Watson, D., & O'Hara, M. W. (2008). Parsing the general and specific components of depression and anxiety with bifactor modeling. *Depression and Anxiety, 25,* E34–E46. http://dx.doi.org/10.1002/da.20432

Sinharay, S., Haberman, S., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice, 26,* 21–28. http://dx.doi.org/10.1111/j.1745-3992.2007.00105.x

Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice, 30,* 29–40. http://dx.doi.org/10.1111/j.1745-3992.2011.00208.x

Steiger, J. H., & Schönemann, P. H. (1978). A history of factor indeterminacy. In S. Shye (Ed.), *Theory construction and data analysis in the behavioral sciences* (pp. 136–178). San Francisco, CA: Jossey-Bass.

Sterba, S. K. (2011). Implications of parcel-allocation variability for comparing fit of item-solutions and parcel-solutions. *Structural Equation*

Modeling, 18, 554–577. http://dx.doi.org/10.1080/10705511.2011.607073

Sterba, S. K., & MacCallum, R. C. (2010). Variability in parameter estimates and model fit across repeated allocations of items to parcels. *Multivariate Behavioral Research, 45,* 322–358. http://dx.doi.org/10.1080/00273171003680302

Stucky, B. D., & Edelen, M. O. (2014). Using hierarchical IRT models to create unidimensional measures from multidimensional data. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 183–206). New York, NY: Routledge/Taylor & Francis Group.

Stucky, B. D., Thissen, D., & Edelen, M. O. (2013). Using logistic approximations of marginal trace lines to develop short assessments. *Applied Psychological Measurement, 37,* 41–57. http://dx.doi.org/10.1177/0146621612462759

Ten Berge, J. M., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika, 69,* 613–625. http://dx.doi.org/10.1007/BF02289858

Wood, R., Kandola, P., Wilson, D., Gibbons, R., Schilling, S., Muraki, E., & Bock, D. (2003). TESTFACT for Windows 4.0.2 [Software]. Lincolnwood, IL: Scientific Software.

Yang, F. M., & Jones, R. N. (2008). Measurement differences in depression: Chronic health-related and sociodemographic effects in older Americans. *Psychosomatic Medicine, 70,* 993–1004. http://dx.doi.org/10.1097/PSY.0b013e31818ce4fa

Yang, Y., & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment, 29,* 377–392. http://dx.doi.org/10.1177/0734282911406668

Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64,* 213–249. http://dx.doi.org/10.1007/BF02294536

Zinbarg, R. E., Barlow, D. H., & Brown, T. A. (1997). Hierarchical structure and general factor saturation of the Anxiety Sensitivity Index: Evidence and implications. *Psychological Assessment, 9,* 277–284. http://dx.doi.org/10.1037/1040-3590.9.3.277

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. http://dx.doi.org/10.1007/s11336-003-0974-7

Zinbarg, R., 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. http://dx.doi.org/10.1177/0146621605278814

Appendix

Working With Real Data

For our demonstrations, we used only the standardized bifactor loadings from a published article. To show how researchers can compute the indices presented herein on their own data, or when working from a reproduced correlation matrix derived from published research, we illustrate the use of readily available software R 3.1.1 (R Development Core Team, 2014) *psych* library version 1.3.2 (Revelle, 2014). To compute the various statistical indices in this example, first, the correlation matrix was reproduced using simulated standardized factor loadings. Fifteen items were simulated to have standardized factor loadings of .6 on the general factor. Five items were specified to have strong (standardized) factor loadings on the first group factor ($\lambda s = .70$), five with moderate factor loadings ($\lambda s = .50$) on the second group factor, and five with weak loadings ($\lambda s = .30$) on the third group factor.

From this loading matrix, the model reproduced correlation matrix was computed using the following transformation: $\hat{\boldsymbol{R}} = \lambda\phi\lambda' + \psi$ where λ is a matrix of standardized factor loadings (15×4), ϕ is a matrix of the intercorrelations among factors (4×4), with unities on the diagonal and zeroes in the off-diagonals, and ψ is a 15×15 matrix with unique variances on the diagonal, zeros elsewhere. The resulting product is the 15×15 reproduced correlation matrix ($\hat{\boldsymbol{R}}$), which, for present purposes, is treated as a proxy for the actual data analyzed. After obtaining the model reproduced correlation matrix, computing the various indices easily can be accomplished.

Using the *omega* (or *omegaSEM*) function built into the *psych* package (Revelle, 2014), the user need only specify two basic elements, a correlation matrix and the number of group factors. The command is: omega(m, nfactors = 3) where *m* is the correlation matrix and *nfactors* is the number of group factors. This requires the user to name the correlation matrix, in this case, *model* (abbreviated "*m*"), and specify the number of group factors to be

(*Appendix continues*)

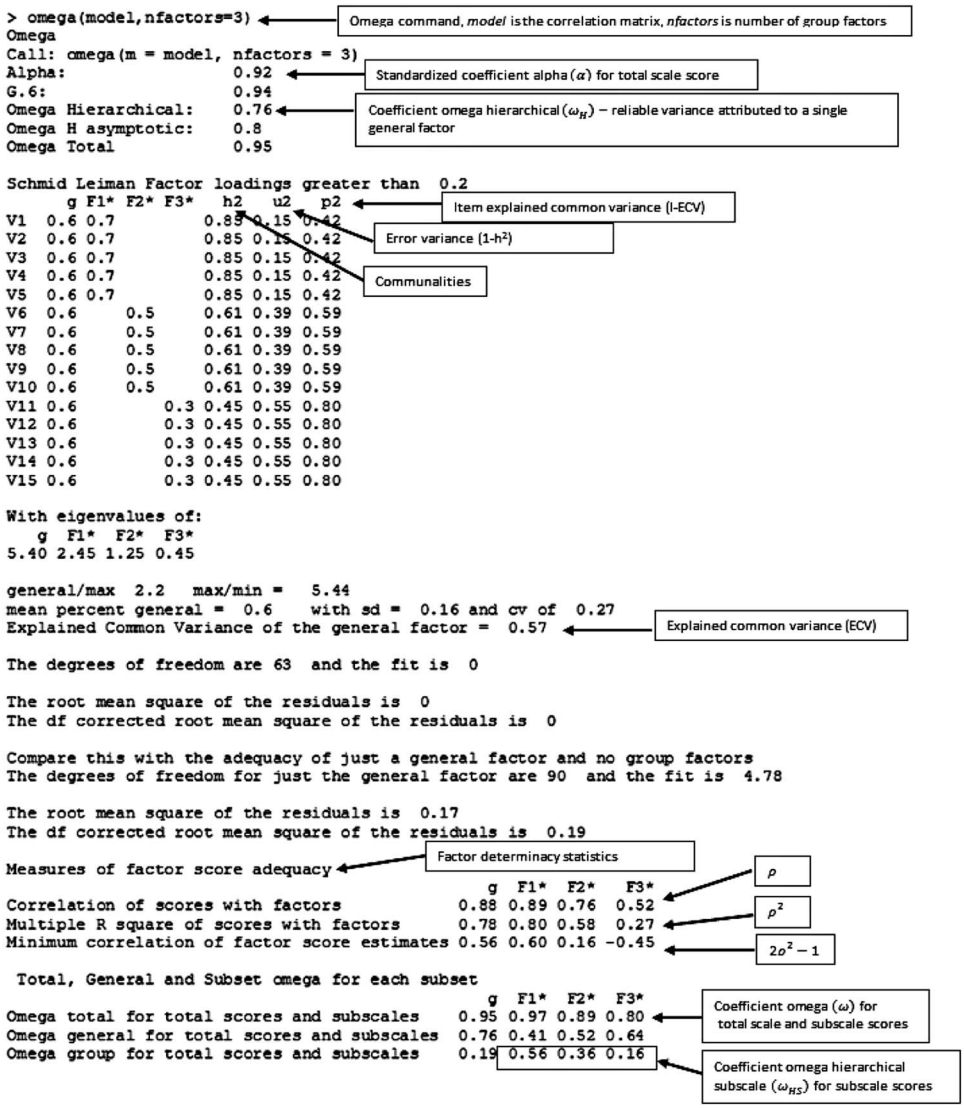

Figure A1. Annotated R output for real data example.

analyzed. That said, it should be noted that there are a variety of other options available in terms of model specification, such as factor methods (e.g., minimum residuals, maximum likelihood), type of correlation matrix being used (e.g., polychoric/tetrachoric, Pearson), number of observations for computing goodness of fit statistics, as well as many others. Annotated results are shown in Figure A1. The only index not available on the annotated output is *H*, which easily can be computed by hand because *H* is a function of the ratio of the items' squared loading (proportion of variance explained by the latent variable) on a factor to 1 minus the squared loading (proportion of variance unexplained by the latent variable) on that factor (see Equation 9).

Received November 7, 2014
Revision received April 1, 2015
Accepted May 17, 2015 ■