Educational Measurement: Issues and Practice
Winter 2015, Vol. 34, No. 4, pp. 14–20

Evaluation of Dimensionality in the Assessment of Internal Consistency Reliability: Coefficient Alpha and Omega Coefficients

Samuel B. Green, *Arizona State University*, and Yanyun Yang, *Florida State University*

In the lead article, Davenport, Davison, Liou, & Love demonstrate the relationship among homogeneity, internal consistency, and coefficient alpha, and also distinguish among them. These distinctions are important because too often coefficient alpha—a reliability coefficient—is interpreted as an index of homogeneity or internal consistency. We argue that factor analysis should be conducted before calculating internal consistency estimates of reliability. If factor analysis indicates the assumptions underlying coefficient alpha are met, then it can be reported as a reliability coefficient. However, to the extent that items are multidimensional, alternative internal consistency reliability coefficients should be computed based on the parameter estimates of the factor model. Assuming a bifactor model evidenced good fit, and the measure was designed to assess a single construct, omega hierarchical—the proportion of variance of the total scores due to the general factor—should be presented. Omega—the proportion of variance of the total scores due to all factors—also should be reported in that it represents a more traditional view of reliability, although it is computed within a factor analytic framework. By presenting both these coefficients and potentially other omega coefficients, the reliability results are less likely to be misinterpreted.

Keywords: coefficient α, internal consistency, reliability, unidimensionality, validity

In the lead article, Davenport, Davison, Liou, & Love (abbreviated as DDLL), described how reliability, dimensionality, and internal consistency are distinct, but related concepts. Their presentation focused on coefficient alpha as an estimate of reliability and built on points made by Cronbach (1951) in his classic article on alpha. Taking this perspective, they offered a number of insights that have not been previously made (e.g., see Equations 7 and 8 in DDLL's article). Overall, we agree with the major premise of their article and have made similar arguments previously in the literature (e.g., Green, Lissitz, & Mulaik, 1977; Green & Yang, 2009a), as have others (e.g., Cortina, 1993; Schmitt, 1996; Sijtsma, 2009). We continue to encounter misinterpretations of coefficient alpha as an index of internal consistency or of homogeneity in the educational and psychological literature, and thus the message delivered by DDLL is needed.

Although we agree with the major points made by DDLL, we choose to view coefficient alpha within a modeling framework. Our preference is to model the relationships among items and choose an internal consistency estimate of reliability that is consistent with the chosen model and the intended purpose of the measure. We recognize that in some sense we are going outside the boundaries set by DDLL, who viewed coefficient alpha through the lens of Cronbach (1951). However, researchers who use a modeling approach are less likely to misapply coefficient alpha and should gain a deeper understanding of the psychometric quality of their test scores. As with any modeling process, researchers can encounter uncertainty in making choices between models, but this uncertainty is a desirable feature if they acknowledge the difficulties that are encountered in their assessment of model fit and the limits of their knowledge.

In our commentary, we give formulas for computing reliability coefficients based on parameter estimates from a factor analytic model. These estimates can be determined using either confirmatory factor analysis (CFA) or exploratory factor analysis (EFA). We then present an example to illustrate the basic points made by DDLL and the advantages of viewing reliability within a modeling process.

Internal Consistency Reliability Estimates Using Factor Analysis

In this section, we present a general formulation for reliability within a factor analytic framework. We then consider special cases and, in particular, reliability when the assumptions of coefficient alpha are met and reliability when a bifactor model is deemed appropriate.

General Factor Analytic Formulation

Within classical test theory (CTT), a score on the j^{th} item (x_j) is the sum of an item true score and an item error score:

$$x_j = t_j + e_j. \tag{1}$$

Samuel B. Green, 851 Cady Mall, P.O. Box 873701, Tempe, AZ 85287-3701, samgreen@asu.edu, Yanyun Yang, Florida State University, 1114 West Call Street, Tallahassee, FL 32306; yyang3@fsu.edu

Alternatively, within a factor analytic model, a score on an item can be expressed as a function of factor(s) and a residual:

$$x_j = a_j + \sum_{m=1}^{M} \lambda_{jm} f_m + r_j, \tag{2}$$

where a_j is the intercept for item j, f_m is a score on the m^{th} factor, λ_{jm} is the loading on the m^{th} factor for item j, and r_j is the residual score for item j. The factor model can be recast as a CTT model if we conceptualize $t_j = a_j + \sum_{m=1}^{M} \lambda_{jm} f_m$ and $e_j = r_j$.

The focus of reliability typically is on scale scores (X), which may be computed by summing across the J item scores:

$$X = \sum_{j=1}^{J} x_j. \tag{3}$$

Within CTT, scale scores are the sum of the scale true scores and the scale error scores:

$$X = T + E, \tag{4}$$

where T is the sum of the item true scores, and E is the sum of the item error scores. Given independence between true and error scores, $\sigma^2(X) = \sigma^2(T) + \sigma^2(E)$, and scale reliability is the ratio of the scale true score variance to the scale score variance,

$$\rho_{XX'} = \frac{\sigma^2(T)}{\sigma^2(X)}. \tag{5}$$

We now can recast reliability within a factor model. Scale scores are

$$X = T + E = \sum_{j=1}^{J} \left(a_j + \sum_{m=1}^{M} \lambda_{jm} f_m \right) + \sum_{j=1}^{J} r_j \tag{6}$$

and reliability is

$$\rho_{XX'} = \frac{\sigma^2(T)}{\sigma^2(X)} = \frac{\sigma^2 \left(\sum_{j=1}^{J} \sum_{m=1}^{M} \lambda_{jm} f_m \right)}{\sigma^2(X)}. \tag{7}$$

A sample reliability coefficient can be computed by substituting sample estimates for the numerator and denominator, which can be determined from the parameter estimates of the fitted model. Alternatively, the numerator could be computed in the same manner, but the denominator could be calculated directly as the variance of the summed item scores. To the extent that a model fits the data, these two methods should yield similar results. A reliability coefficient defined in this way is referred to in the literature as coefficient omega (McDonald, 1999).

Coefficient Alpha and the Factor Analytic Model

An assumption underlying coefficient alpha is essential tau equivalence; that is, a single factor underlies all items with a common factor loading, although intercepts can vary across items. Additional assumptions are that the errors for any item must have a mean of zero, be uncorrelated with factor scores, and be uncorrelated with the errors for any other item (the latter being referred to as the uncorrelated errors assumption). Given these assumptions, alpha is equal to reliability and is a function of the mean covariance between items, $\bar{\sigma}_{jj'}$; more precisely,

$$\rho_{XX'} = \alpha = \frac{J^2 \bar{\sigma}_{jj'}}{\sigma^2(X)}. \tag{8}$$

Given a factor model that is consistent with the assumptions underlying coefficient alpha, Equation 2 for an item simplifies to

$$x_j = a_j + \lambda f + r_j. \tag{9}$$

Fixing the variance of the factor at 1 to define its metric, Equation 7 for reliability can be reexpressed as

$$\rho_{XX'} = \frac{\sigma^2(T)}{\sigma^2(X)} = \frac{\sigma^2 \left(\sum_{j=1}^{J} \lambda f \right)}{\sigma^2(X)} = \frac{\left(\sum_{j=1}^{J} \sum_{j=1}^{J} \lambda^2 \right)}{\sigma^2(X)} = \frac{J^2 \lambda^2}{\sigma^2(X)}. \tag{10}$$

In words, if the model for coefficient alpha is correct, the item true score covariance matrix has variances of λ^2 for all items and covariances of λ^2 for all pairs of items in the population. Because $\sigma^2(T)$ is the sum of all elements in this matrix, $\sigma^2(T) = J^2\lambda^2$. The implication also is that $J^2\lambda^2 = J^2\bar{\sigma}_{jj'}$, and $\rho_{XX'} = \alpha$.

Reliability and the Bifactor Model

The model underlying coefficient alpha is unnecessarily restrictive for a couple of reasons. From a modeling perspective, items on a measure are unlikely to be unidimensional (and even less likely to have equivalent factor loadings). Thus, when modeling item data, we generally would want to assess the fit of less constrained, multidimensional models. From an application viewpoint, a measure does not need to be unidimensional to yield interpretable scores. Rather a measure just needs to be sufficiently unidimensional so that individuals who score high (or at some other level L) on this measure are likely to be at a high level (or at level L) on its primary, underlying dimension and not because of their levels on other dimensions.

We focus on the bifactor model as an alternative to the unidimensional model underlying coefficient alpha. This model has evidenced good fit to data for items on a variety of measures (e.g., Bados, Gómez-Benito, & Balaguer, 2010; Farias et al., 2008; Gignac, Palmer, & Stough, 2007; Patrick, Hicks, Nichol, & Krueger, 2007; Thomas, 2012). In addition, the bifactor model allows for an assessment of whether items on a measure are sufficiently unidimensional that its scores are interpretable (Reise, 2012).

A bifactor model includes a general factor (f_{GEN}) that underlies all items on a measure and one or more group factors (f_{GRP1}, f_{GRP2}, …) that underlie subset(s) of items:

$$x_j = a_j + \lambda_{j\ GEN} f_{GEN} + \lambda_{j\ GRP1} f_{GRP1} + \lambda_{j\ GRP2} f_{GRP2} + \cdots + r_j. \quad (11)$$

Following Reise (2012), the general factor and the group factors all are orthogonal to each other. Presumably, the general factor represents the broad construct of interest, whereas the group factors represent narrower subdomain constructs.

Given the variances of the factors are fixed at 1, the equation for the reliability of a scale with an underlying bifactor model with G group factors is relatively simple,

$$\rho_{XX'} = \frac{\sigma^2(T)}{\sigma^2(X)} = \frac{\left(\sum_{j=1}^{J} \lambda_{j\ GEN}\right)^2 + \left(\sum_{j=1}^{J} \lambda_{j\ GRP\ 1}\right)^2 + \left(\sum_{j=1}^{J} \lambda_{j\ GRP\ 2}\right)^2 + \cdots + \left(\sum_{j=1}^{J} \lambda_{j\ GRP\ G}\right)^2}{\sigma^2(X)}. \quad (12)$$

It is also referred to as coefficient omega (McDonald, 1999; Reise, 2012; Zinbarg, Revelle, Yovel, & Li, 2005). Each quantity in the numerator on the right side of the equation represents the contribution of a factor to the true score variance of the scale scores (computed by summing across items). Because the factors are orthogonal, the total true score variance of the scale scores is the sum of the true score variance for general and group factors.

Additional omega indices can be computed for measures with a bifactor structure. In particular, an index can be computed to assess the proportion of variance of the scale scores (summed items) due to the general factor,

$$\omega_H = \frac{\left(\sum_{j=1}^{J} \lambda_{j\ GEN}\right)^2}{\sigma^2(X)}. \quad (13)$$

This index is referred to as coefficient omega hierarchical (McDonald, 1999; Reise, 2012; Zinbarg et al., 2005), and allows researchers to state the degree that summed item scores are saturated by the general factor. Higher values on ω_H indicate greater confidence in interpreting the scale scores as due to the general factor and not to group factors (and measurement error). The temptation is to state a cutoff value for ω_H that represents the lower bound for interpretability of scale scores. However, cutoff values for ω_H should depend at least partially on the relative importance of the outcomes associated with the decisions made based on the measure. For example, we should require a measure to have a higher cutoff value on ω_H if it is used to make final, high-stakes decisions rather than for screening decisions.

An omega index also can be computed to assess whether subscale scores for a measure with a bifactor structure should be reported. Coefficient omega for the subscale g (ω_{S_g}) assesses the proportion of variance of the subscale scores (summed items) due to the group factor for that subscale,

Table 1. Covariance (Lower-Left Triangle) and Correlations (Upper-Right Triangle) Between Items

Items	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_1	1.40	.449	.464	.449	.214	.138	.214	.207	.276	.267
x_2	.65	1.50	.483	.467	.207	.067	.138	.133	.200	.194
x_3	.65	.70	1.40	.483	.214	.207	.214	.138	.276	.267
x_4	.65	.70	.70	1.50	.138	.133	.207	.200	.200	.226
x_5	.30	.30	.30	.20	1.40	.449	.464	.449	.276	.301
x_6	.20	.10	.30	.20	.65	1.50	.483	.467	.200	.194
x_7	.30	.20	.30	.30	.65	.70	1.40	.483	.276	.301
x_8	.30	.20	.20	.30	.65	.70	.70	1.50	.200	.258
x_9	.40	.30	.40	.30	.40	.30	.40	.30	1.50	.355
x_{10}	.40	.30	.40	.35	.45	.30	.45	.40	.55	1.60

$$\omega_{S_g} = \frac{\left(\sum_{j=1}^{J_g} \lambda_{j\ GRP_g}\right)^2}{\sigma^2(X_g)} \quad (14)$$

(Reise, 2012). Note that the summation is from the first to the last item on the subscale. Higher values on ω_{S_g} indicate greater confidence in interpreting the scale scores as due to the group factor of interest.

Gignac and Watkins recently (2013) published an article that illustrated the usefulness of omega coefficients with the Wechsler Adult Intelligence Scale (4th edition). An ω_H of .86 on the full scale intelligence quotient indicated that it is sufficiently saturated by the general factor that it can be interpreted. On the other hand, the ω_{S_g}'s on the clinical index scores ranged in value from .13 to .47, suggesting that interpretation of index scores probably is unjustified.

Example

We have created the following fictional example to illustrate interpretation of various reliability coefficients. The data are from 500 participants and represent scores on a scale consisting of 10 items with seven-point response scales. The item variances and covariances for this scale are in Table 1 as well as item correlations.

Given the relatively large number of points on the response scale, the data were treated as continuous (Rhemtulla, Brosseau-Liard, & Savalei, 2012), and linear factor analysis models were fit to the item covariance matrix. Linear factor analysis can yield relatively accurate parameter estimates if the number of ordered categories is five or greater (see Rhemtulla et al., 2012). Green and Yang (2009b; Green et al., in press) offer an alternative approach for estimating reliability coefficients using a nonlinear factor analysis method for items with a more limited number of response categories.

We chose CFA rather than EFA for our example because researchers should have testable models if measures are developed using well-developed blueprints. That said, EFA is the preferred method when researchers are unable to generate models because insufficient information is known about measures to specify models or because models failed to fit when tested with CFA.

Initially, we compute coefficient alpha for our example data and discuss its meaning. Next, we explore the factor structure underlying the measure with the goal of choosing the most appropriate model. Once we have chosen a model, we assess the reliability of the measure using a variety of coefficients. Finally, we consider the relationship between the various internal consistency reliability coefficients and the homogeneity of a measure.

Coefficient Alpha

Substituting sample quantities for parameter values in Equation 8, we computed coefficient alpha for our example: $\hat{\alpha} = J^2 \bar{\hat{\sigma}}_{jj'} / \hat{\sigma}^2(X) = 10^2(.411)/51.7 = .795$. As indicated by DDLL, coefficient alpha is distinct from, but related to internal consistency and homogeneity. Following Cronbach (1951) and to simplify some proofs, DDLL formulated the relationship among these quantities assuming all item variances are equal to 1.0, that is, based on an item correlation matrix. For our example an alpha based on the item correlation matrix (i.e., standardized alpha: $\hat{\alpha}_S$) was equal to .796. The scale's internal consistency, as assessed by the mean correlation between items ($\bar{r}_{ii'}$), was .281, and scale homogeneity, as measured by the proportion of variance accounted for by the first principal component (V), was .354. DDLL offered an approximation to standardized coefficient alpha (Equation 8 in their article) that clarifies the relationship among the three statistics, $\hat{\alpha}_S \approx \bar{r}_{jj'}/V$, assuming the approximation is reasonably accurate. DDLL indicated that the approximation works well, particularly when the item correlations do not differ dramatically from each other; for our example, the approximation yielded a value of .794 (i.e., $\hat{\alpha} \approx .281/.354 = .794$), extremely close to the computed standardized alpha of .796.

If coefficient alpha is relatively robust to violations of its underlying assumptions, then an argument could be made that no further analyses need be conducted, and in this case, an alpha of .795 could be reported as the internal consistency reliability coefficient. Although coefficient alpha can yield an inaccurate estimate of reliability under violation of the essential tau equivalency assumption, it is likely to be relatively robust to violation of this assumption if a test is carefully developed (e.g., Feldt & Qualls, 1996; Green & Yang, 2009a; Raykov, 1997). On the other hand, coefficient alpha is not robust to violation of the uncorrelated errors assumption. Although it is largely accepted that the uncorrelated errors assumption is likely to be violated with speeded tests and with measures containing context-dependent item sets, it is less clear how often this assumption is violated more generally (e.g., Gessaroli & Folske, 2002; Green & Hershberger, 2000; Green & Yang, 2009a; Zimmerman, Zumbo, & Lalonde, 1993). Regardless of the accuracy of coefficient alpha as an estimate of internal consistency reliability, we think that the use of a modeling process is likely to yield more informative reliability coefficients and ones that are less open to misinterpretation, as we illustrate next.

The Modeling Process

Within our example, we presume that the 10-item measure was designed based on a blueprint, and this blueprint implied a series of unidimensional and bifactor models. All models are analyzed using maximum likelihood estimation method. The results for these models are presented in Table 2. We initially fit the model underlying coefficient alpha, as shown in the top part of Figure 1. All items load on a single factor; the loadings on this factor are constrained to be equal across items with a common value of λ; the covariances between errors are fixed to 0; and the item intercepts are allowed to differ (intercepts not shown). Given this model, all model-implied item covariances are equivalent and equal to λ^2, and reliability simplifies to $\rho_{XX'} = J^2 \lambda^2/\sigma^2(X)$ (i.e., Equation 10). As shown in Table 2, reliability for our example is equal to .797, very close to the computed alpha of .795[1]. However, as also shown in the table, the model fits poorly. From a modeling perspective, poor fit implies that we should interpret neither the estimated parameters of the model nor coefficients based on these estimates. Rather, we should evaluate the fit of less constrained, alternative models. Accordingly, we removed the equality constraints on the factor loadings, yielding a congeneric model. As shown in Table 2, the congeneric model also evidenced poor fit.

We then evaluated two bifactor models. With the first bifactor model, all 10 items were functions of a general factor, and in addition the first four items were a function of a group factor. The second model had the same structure as the first bifactor model, but included a second group factor with loadings on the fifth through eighth items, as shown at the bottom of Figure 1. The first bifactor model with a single group factor fit relatively well in comparison with the unidimensional models, but the second bifactor model with two group factors demonstrated even better fit. We then examined the standardized loadings (i.e., correlations given orthogonal factors) for the second bifactor model to ensure that the loading were sufficiently high across items to warrant labeling the factors *general* and *group*. The standardized loadings on the general factor ranged from .337 to .613, and the standardized loadings on the two group factors ranged from .463 to .608. Based on these results, we concluded that the bifactor model with two group factors should be used to assess the reliability of the measure.

We computed $\hat{\omega}$ based on the preferred bifactor model by substituting sample estimates into Equation 12. As shown in Table 2, the reliability estimate is .843; that is, approximately 84% of the variance of the scale scores (computed by summing across the 10 items) is accounted for by the general and the two group factors. This coefficient is approximately .05 greater than the coefficient alpha of .795. Given the preferred bifactor model demonstrated substantially better fit than the model underlying coefficient alpha, we are more comfortable with the reliability estimate of .843. Nevertheless, we, as researchers, may still feel some pangs of anxiety in our choice of a reliability coefficient. Reliabilities based on bifactor models are likely to increase as one specifies additional group factors. Thus, researchers should be careful not to capitalize on the chance characteristics of a sample by adding spurious group factors and, in so doing, increasing the estimated reliability. In the context of our example, we are more comfortable with interpreting the coefficient of .843 if we hypothesized the two group factors *a priori* rather than arriving at this model through a specification search.

Table 2. Reliabilities and Fit Indices for Example with N = 500

	Reliability Analysis			Model Fit			
Model	**True Score Variance**	**Error Score Variance**	**Reliability**	**χ^2**	**df**	**CFI**	**RMSEA**
Essential tau equivalent	41.518	10.579	.797	436.669	44	.686	.134
Congeneric	41.158	10.569	.796	431.458	35	.683	.151
Bifactor with 1 group factor	41.421	8.803	.825	101.949	31	.943	.068
Bifactor with 2 group factor	43.521	8.107	.843	29.743	27	.998	.014

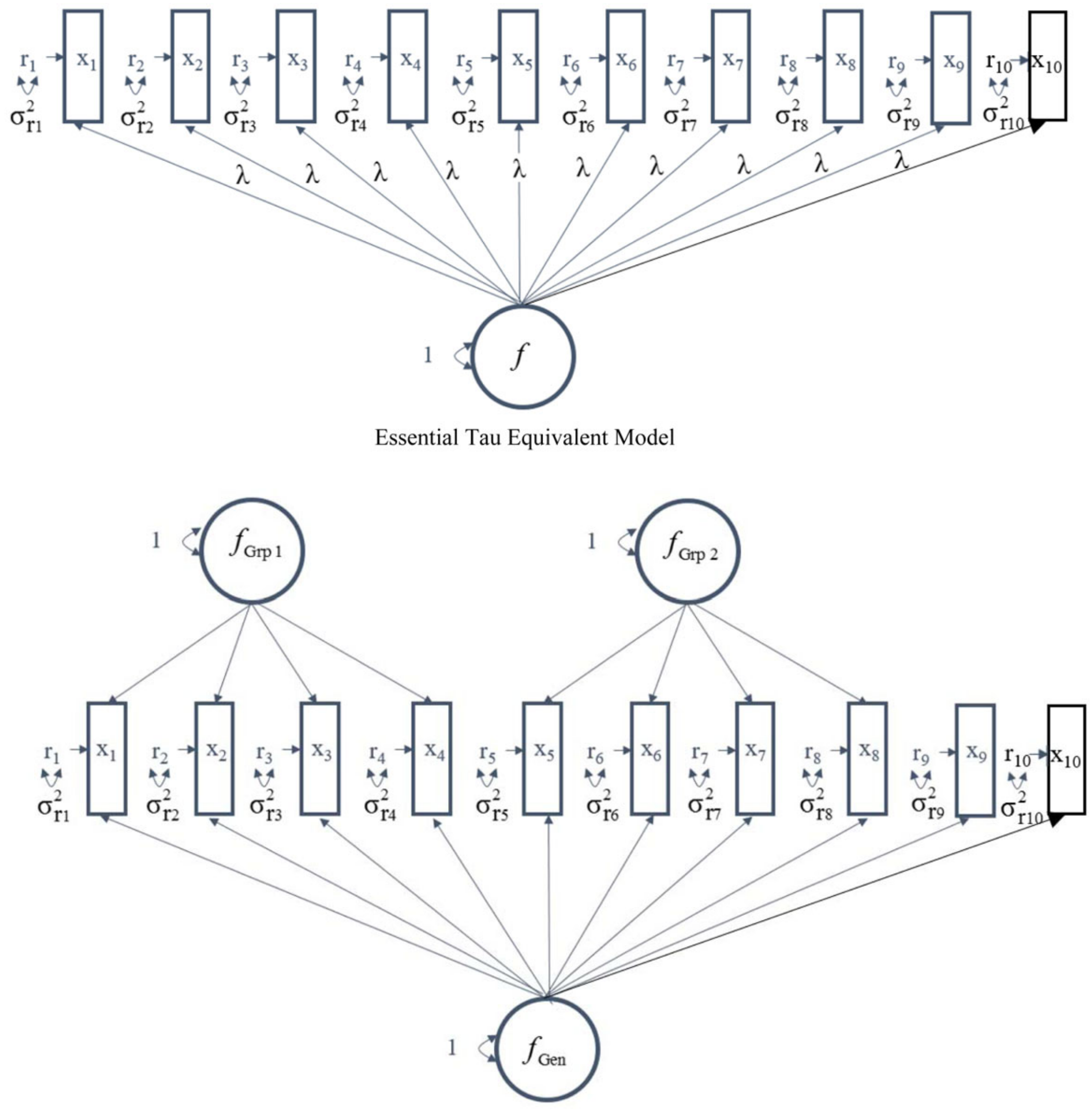

FIGURE 1. Two models fit to the covariance matrix in Table 1 for the example.

Table 3. Variances and Reliabilities for Various Summed Observed Scores Based on Bifactor Model

	Variances of					Reliability Based on		
Summed Observed Scores	**General Factor**	**Group Factor for x_1 to x_4**	**Group Factor for x_5 to x_8**	**Error**	**Total**	**All Factors**	**General factor only**	**Group factor(s) Only**
Sum of x_1 to x_{10}	29.690	7.120	6.711	8.117	51.628	.843	.575	.268
Sum of x_1 to x_4	3.716	7.120	0	3.064	13.900	.779	.267	.512
Sum of x_5 to x_8	4.140	0	6.711	3.050	13.900	.781	.298	.483
Sum of x_9 to x_{10}	2.210	0	0	1.993	4.203	.526	.526	.000

Reliability Coefficients Based on the Preferred Bifactor Model

In Table 3, we present a variety of coefficients based on the preferred bifactor model that allow for a deeper understanding of the measure. We have some apprehension about the interpretation of overall scale scores if the objective is to assess individuals on the general factor. Although, as previously discussed, the reliability of the overall scale ($\hat{\omega}$) is .843, the proportion of scale variance due to the general factor ($\hat{\omega}_H$) is just .575. The additional proportion of reliable variance of .268 is due to the group factors. Provocatively, we could compute scores based on only two of the items, and the proportion of variance for these scores due to the general factor of .526 would be almost as great. On the surface, it would seem preferable to develop a few more items like items 9 and 10 and combine them together to create an overall scale rather than use the 10-item scale. Of course, before making this type of decision, we would have to consider the meaningfulness of the scores from a broader psychometric perspective.

The omega coefficients also allow us to assess the quality of potential subscale scores. In our example, we could sum items 1 through 4 to obtain scores for one subscale and sum items 5 through 8 to obtain scores for the second subscale. For simplicity, we will focus on the first subscale and assume that an objective of the measures was to assess where individuals fall on the first group factor. In this instance, the proportion of subscale variance due to the first group factor ($\hat{\omega}_{S_1}$) was .512. Approximately a quarter of variance of the subscale scores was due to the general factor. It is quite possible that items cannot be written that only assess the first group factor. However, if additional items like the first four items were included on the subscale, $\hat{\omega}_{S_1}$ might reach a more acceptable level.

When we interpreted the results in Table 3, we assumed that all factors were of interest. In practice, this scenario may not hold. An alternative context might be that the general factor is a methodological artifact (e.g., social desirability or halo effect), and the interest is on the group factors. Within this context, we would like the proportion of variance of the various summed scores due to the general factor to be as small as possible.

Internal Consistency Reliability Coefficients, Homogeneity, and Validity

Consistent with DDLL, we view homogeneity as the degree to which items assess a single dimension. We will consider a few indices to assess degree of unidimensionality, although others are available (e.g., the DETECT index by Zhang and Stout, 1999). Following Cronbach (1951), DDLL operationalized degree of unidimensionality as the proportion of variance accounted for by the first principal component (V) and showed that V is distinct from, but related to coefficient alpha. In contrast, Reise (2012) discussed *explained common variance* (ECV) as a straightforward index to assess unidimensionality within a factor analytic framework; ECV is the proportion of common variance (based on all common factors) associated with the first common factor. Unfortunately the values of V and ECV for our data of .354 and of .663, respectively, are not very informative to test users. One can define a cutoff value to delineate a "satisfactory degree" of unidimensionality (e.g., greater than .20 for V, as discussed by DDLL), but a cutoff does not solve the more basic problem: the indices fail to tell us about the unidimensionality of the scores that are interpreted by test users. As argued by Reise (2012), $\hat{\omega}_H$ is a useful index of unidimensionality because it focuses on the scores that are frequently interpreted by test users, that is, scale scores computed by summing across items. In the context of our example, $\hat{\omega}_H = .575$, which means that 57.5%, of the variance of the scale score variance is due to a single, general dimension.

We believe that $\hat{\omega}_H$ is the preferred index of unidimensionality if the scores of interest are summed item scores and a bifactor model is conceptually meaningful and fits the data; however, researchers must still face the difficult decision about what value of $\hat{\omega}_H$ indicates sufficient unidimensionality to conclude that the scale scores are interpretable. What constitutes sufficient unidimensionality cannot be made only on the value of $\hat{\omega}_H$, but rather on a broad assessment of validity. We would want to examine evidence concerning whether the group factors are conceptually meaningful or methodological artifacts that can be avoided if the measure is designed carefully. Additional information would be helpful, such as the quality of alternative measures designed to measure the construct of interest and whether the scores are used to make high-stakes decisions.

It is important to note that the coefficient $\hat{\omega}_H$ blurs the boundaries between reliability and validity coefficients (McDonald, 1999). With $\hat{\omega}_H$, we are interested in the relative contribution of the general factor. In contrast, with reliability coefficients (including $\hat{\omega}$), we are interested in the relative contribution of all factors, group factor(s) as well as the general factor. Within the framework of validity, additional sources of evidence are required to adequately interpret $\hat{\omega}_H$.

Conclusion

Coefficient alpha can provide a relatively accurate assessment of reliability under some conditions, but not under others. Regardless, researchers may misinterpret it as an index of internal consistency or homogeneity. The various omega coefficients provide more detailed information about the reliability of measures and are less likely to be misinterpreted because of the transparent relationship between these coefficients and the factor model that they are based on. Coefficient omega, which provides an overall assessment of reliability, can be biased if its underlying model is misspecified. On the other hand, the researcher is required to discuss the model underlying coefficient omega rather than just assume a highly restrictive one, as with coefficient alpha. Also, additional omega coefficients can be computed to allow researchers to assess the contribution of general and group factors to the reliability of scales (given a bifactor model provides good fit). For example, coefficient omega hierarchical is the proportion of scale variance due to the general factor. Researchers can differentiate between reliability and homogeneity by reporting coefficient omega and coefficient omega hierarchical, respectively.

As with any set of statistical indices, omega coefficients should come with a warning label. Omega coefficients are internal consistency estimates of reliability and thus are unaffected by some types of measurement errors. For example, they are insensitive to transient errors in that they are based on the administration of a measure on a single occasion. Also, omega coefficients are reliability indices of summed item scores and not scale scores that are nonlinear transformations of summed item scores. In addition, they require modeling of item data, and thus researchers must have the skills to work with the complexities inherent with factor analytic models. For the same reason, omega coefficients require studies with large samples. Thus, if researchers wish to assess the reliability of measures and have a limited number of research participants, other coefficients must be sought. One possibility is to compute coefficient alpha based on items combined into parcel to better meet the assumptions. As discussed by LLDD and others (Green & Yang, 2005; Green et al., in press), parcels must be carefully constructed to avoid inaccurate reliability estimates using this approach.

Note

[1] If parameters are estimated using unweighted least squares, the model-implied covariances are equal to the mean covariance among items because the mean minimizes the sum of squared residuals. Thus, if the model is estimated using unweighted least squares, the reliability based on CFA is equivalent to coefficient alpha, $\hat{\rho}_{XX'} = \hat{\alpha} = J^2 \bar{\hat{\sigma}}_{jj'} / \hat{\sigma}^2(X)$ (Miller, 1995).

References

Bados, A., Gómez-Benito, J., & Balaguer, G. (2010). The state-trait anxiety inventory, trait version: Does it really measure anxiety? *Journal of Personality Assessment*, *92*, 560–567.

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

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

Davenport, E. C., Davison, M. L., Liou, P.-Y., & Love, Q. U. (2015). Reliability, dimensionality, and internal consistency as defined by Cronbach: Distinct albeit related concepts. *Educational Measurement: Issues and Practice*, *34*, 1–6.

Farias, S. T., Mungas, D., Reed, B. R., Cahn-Weiner, D., Jagust, W., Baynes, K., & DeCarli, C. (2008). The measurement of everyday cognition (ECog): Scale development and psychometric properties. *Neuropsychology*, *22*, 531–544.

Feldt, L. S., & Qualls, A. L. (1996). Bias in coefficient alpha arising from heterogeneity. *Applied Measurement in Education*, *9*, 277–286.

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling*, *16*, 625–641.

Gessaroli, M. E., & Folske, J. C. (2002). Generalizing the reliability of tests comprised of testlets. *International Journal of Testing*, *2*(3–4), 277–295.

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.

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.

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.

Green, S. B., Lissitz, R. W., & Mulaik, S. (1977). Limitations of coefficient alpha as an index of text unidimensionality. *Educational and Psychological Measurement*, *37*, 827–839.

Green, S. B., & Yang, Y. (2005, April). *K-Split coefficient alpha*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

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

Green, S. B., & Yang, Y. (2009b). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika*, *74*, 155–167.

Green, S. B., Yang, Y., Alt, M., Brinkley, S., Gray, S, Hogan, T., & Cowan, N., (in press). Use of internal consistency coefficients for estimating reliability of experimental task scores. *Psychonomic Bulletin & Review*.

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

Patrick, C. J., Hicks, B. M., Nichol, P. E., & Krueger, R. F. (2007). A bifactor approach to modeling the structure of the Psychopathy Checklist—Revised. *Journal of Personality Disorders*, *21*, 118–141.

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

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

Rhemtulla, M., Brosseau-Liard, P. E., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*, 354–373.

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

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

Thomas, M. L. (2012). Rewards of bridging the divide between measurement and clinical theory: Demonstration of a bifactor model for the Brief Symptom Inventory. *Psychological Assessment*, *24*, 101–113.

Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika*, *64*, 213–249.

Zimmerman, D. W., Zumbo, B. D., & Lalonde, C. (1993). Coefficient alpha as an estimate of test reliability under violation of two assumptions. *Educational and Psychological Measurement*, *53*, 33–49.

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.

