Review of Psychology,
2014, Vol. 21, No. 2, 111-121 ORIGINAL SCIENTIFIC PAPER UDC 159.9

Estimating reliability coefficient for multidimensional measures: A pedagogical illustration

WAHYU WIDHIARSO and HAMDOLLAH RAVAND

The literature has shown that the assumption of unidimensional measurement in psychology is difficult to fulfill, since measurement in psychology is usually a complex process that aims at providing information about constructs. As a consequence, factor analysis for psychological measurement tends to conceal several factors that underlie the items on the scale. Since applying a reliability coefficient (e.g., Cronbach's alpha) based on a unidimensional assumption for a multidimensional measure will underestimate reliability, researchers should use an appropriate coefficient that matches the characteristics of the measure. There are several, albeit not frequently utilized reliability coefficients for multidimensional measures. The present article demonstrates the application of the stratified alpha, Mosier's, Raykov's, McDonald's, and Hancock-Mueller's coefficients for estimating the reliability of multidimensional measures.

Key words: reliability coefficient, multidimensional measurement

The development of assessment instruments in psychology has the central goal of assessing the general construct of interest, which is assumed to be a single construct. Such a measure is termed a *unidimensional measure*. The dimension of the measure is important for interpreting the obtained score. Unidimensional measures require only simple interpretation, since all items on the scale represent a single attribute. In contrast, multidimensional measures require complex interpretation. Therefore, the literature suggests that, in the first step of analysis, scale developers should run a factor analysis to determine the dimensionality of their measure, before examining reliability (Armor, 1974).

Several studies have shown that the assumption of unidimensionality is likely difficult to be fulfilled since factor analysis tends to produce new emerging factors that contribute to the explained scores variance. Some authors believe that it is hard to locate a single factor when measuring a broad ability or trait, especially when the measure contains several items with different levels of precision (Kamata, Turhan, & Darandari, 2003). A single factor can be achieved only when items on a scale are homogenous: this means that items measure similar content with small number of items, focus on narrowly defined content, and have the same level of precision (Graham, 2006). Scales developed in psychology usually measures broad constructs that cover various aspects of trait manifestation. To achieve this purpose, scale developer increases the number of items to expand the domain being measured. Ideally, measures are assigned on the basis of a single dimension of individual differences, but in practice researchers often deal with clusters of items that represent different constructs. Therefore, measurement in psychology is a complex process that aims at providing information about constructs or aspects of behavior (Raykov & Shrout, 2002).

There are several factors that cause scales to be multidimensional. First, constructs in psychology tend to be multidimensional rather than unidimensional in nature (Drolet & Morrison, 2001). In contrast to the natural sciences, which usually deal with observable constructs that can be measured using a single instrument (e.g., measuring length with a ruler), variables in psychology are non-observable (latent), and can be measured using several facets or indicators. For example, to measure self-esteem, DeVellis (2011) suggests that scale developers utilize several items to represent the various indicators of self-esteem. However, using many items can have adverse consequences, as each item might potentially measure another attribute, depending on the scale and the level of *precision of measurement*.

Wahyu Widhiarso, Faculty of Psychology, Universitas Gadjah Mada, Jl. Humaniora No 1 Yogyakarta, Indonesia. E-mail: wahyu_psy@ugm.ac.id (the address for correspondence);

Hamdollah Ravand, Department of English, Faculty of Humanities, Vali-Asr University of Rafsanjan, Iran.

Scales usually consist of several items differing in the size of correlation with the attribute, which in turn can contribute to the emergence of new factors. These items generally should be removed from the scale, but the situation becomes problematic when the content of items contributes substantially to the assessment of the attribute being measured. In such a situation, a scale can be considered a multidimensional rather than a unidimensional measure.

Sometimes the emergence of new factors is attributable to measurement administration rather than the content. Crocker and Algina (1986) gave an example of an aptitude test administered under strict time limits. The score of the test is potentially affected by the irrelevant variable (here, speed), which reflects another attribute than the construct being measured. In this situation, applying a factor analysis will produce several factors, one of them related to an individual's ability to function under time pressure.

Finally, the number of items on a scale can change a unidimensional measure into a multidimensional one. Drolet and Morrison (2001) showed that number of items on a scale can affect the factor structure of the measure. A larger number of items tend to generate a greater potential for factor analysis to yield multiple dimensions, in addition to the original dimension intended to be measured.

Empirical results have shown that the assumption of the unidimensionality of measures usually does not seem to hold (Brunner & Süβ, 2005). On a practical level, researchers usually employ Cronbach's alpha coefficient to estimate reliability in an arbitrary way, even though the unidimensionality assumption is violated (Schmitt, 1996). However, alpha for a scale comprising several dimensions will generally underestimate true reliability (Gerbing & Anderson, 1988), and is therefore suited only for estimating the reliability for a set of items that assess a single latent construct (i.e., unidimensional scales). Thus, when applying an alpha coefficient to a multidimensional measure, researchers are advised to separately estimate reliability for the composite dimensions; if, for example, a scale comprises five dimensions, then the reliability should be estimated in five separate alpha coefficients. For example, a five-factor personality scale is a single scale that measures five separate dimensions. The alpha coefficient then is employed to estimate the five dimensions separately, as using a single alpha to estimate all items on the entire scale would increase the bias estimation of the true reliability. Alternatively, instead of estimating reliability separately for each factor, researchers can apply a multidimensional reliability coefficient, which can incorporate all scale factors into a single measure of reliability.

Multidimensional Reliability in Various Studies

The multidimensionality of measurement is evidenced by the results of factor analyses that generate multiple factors, or for which inter-correlation among items remains zero after controlling for the factors. The multidimensional approach was introduced by Thurstone (1931) and promoted by Burt (1938), employed as an alternate conceptualization of multidimensional measures. When factor analysis for a set of items generates multiple dimensions, one possible way to specify the structure is by defining new groupings of items as sub-attributes. Hence, a multidimensional measure basically consists of several unidimensional measures.

The dimensionality of measures sometimes does not merely refer to how many factors are included in the measure; in addition to being confirmed by the number of factors inside the measure, dimensionality is also confirmed by whether items on the scale support the congeneric measurement assumption. Congeneric measurement is indicated when a set of items on the scale have different factor loadings and error variances. Parallel model is a special case of the congeneric model. Factor loading shows the relationship between items and constructs; thus, if congeneric assumptions are to hold, items on the scale measure the construct differently. When congeneric assumption is held, some items might load strongly on the construct being measured, while others load only moderately.

According to this definition, multidimensional measures comprise heterogeneous items (McDonald, 1999), and unidimensional measures comprise homogeneous items (Carmines & McIver, 1981; Jöreskog, 1971). Using this knowledge, the multidimensionality of a measure can be inferred from either the item bundle level (each set of items assesses a different domain measure) or from the item level (each item has varying factor loading value). The term *multidimensional*, as used in this article, refers to the first condition, whereby a set of items (e.g., subscale, item parcel) measures a sub-attribute.

The most popular method of estimating reliability was developed under the unidimensionality assumption (i.e., coefficient alpha). Hence, one should apply item analysis to each dimension. In addition to estimating the reliability of each dimension separately, one could estimate the reliability of different items together. Forcing a unidimensional measure approach onto multidimensional measures should be avoided, as this will lead to bias. This article will describe how to compute several coefficients that are appropriate for multidimensional measures.

Many personality scales comprise several components, aspects, or facets of the attribute being measured. For example, Big-Five Personality Inventory developed under five-factor personality theory comprises five subscales, each representing one facet of personality: extroversion, neuroticism, conscientiousness, agreeableness, and openness to experience (McCrae & Costa, 1996). Learning Style Inventory comprises four subscales: diverging, assimilating, converging, and accommodating. Coping Strategies Inventory (Lazarus, 1991) reflects two essential styles of coping: problem-focused coping and emotion-focused coping. Intelligence test by Brunner & Süβ (2005) consists of eight

factors: mental speed, memory, reasoning, creativity, figural ability, verbal ability, and numerical ability. Such factors or dimensions can be defined as subscales, subtests, clusters, or testlets (Rae, 2007).

There is only scant literature explaining the using of reliability coefficients for multidimensional measures. One example is a study by Olson and colleagues (2011), who used a multidimensional reliability coefficient when reporting the psychometric properties of their measure, employing a stratified alpha coefficient to estimate the reliability of the Adaptive Capacity Index. In addition, they also employed an internal consistency coefficient (i.e., alpha) for each of their four subscales. Hendriks, Kuyper, Lubbers, and Van der Werf (2011) also employed a stratified alpha coefficient for estimating reliability of the Five-Factor Personality Inventory (FFPI). Another study, conducted by Schretlen, Benedict, and Bobholz (1994), also employed a composite reliability measure for the Wechsler Adult Intelligence Scale, and Evans (1996) conducted a study that used the Mosier composite reliability measure. Finn, Sawyer, and Behnke (2009) also used the Mosier composite reliability coefficient to estimate the reliability of the Psychological State Anxiety Scale. Based on these studies, multidimensional reliability testing can clearly be employed for assessments of a broad range of psychological attributes. Researchers who want to estimate reliability in multidimensional measures are advised to use one of the reliability coefficients that accommodate multidimensional measures, discussed in the upcoming section.

Reliability Coefficients for Multidimensional Measures

Reliability coefficients described in this section are derived from two approaches: reliability coefficient based on classical test theory (CTT; e.g., stratified alpha coefficient) and latent trait (common factor) theory which is implemented with factor analytic approach (e.g., omega coefficient). Several views, which are promoted by authority like McDonald (1999), posit that the estimates derived from these theories are not very different. This notion is supported by the fact that CTT is developed from common factor theory which can be traced to Charles Spearman's concept who described how to recognize that tests measure a common factor and determine the amount of error in test scores (Bovaird & Embretson, 2008). As a consequence, the estimates yielded from both theories are comparable. McDonald (1999) mentioned omega coefficient to be considered an alternative to alpha coefficient. Alpha coefficient is lower bound to omega coefficient. They are equal if and only if the items fit the single-factor model with equal factor loading. Factor loadings can be used to assess item specific as well as scale specific reliability, thus, factor analytic approach is consistent with CTT. A work conducted by Kamata et al. (2003) comparing three multidimensional reliability coefficients based on different theories (CTT vs. common factor) indicates that the estimates from both theories are comparable.

In contrast to the above views, several authors (i.e., Borsboom, 2005) argue that there is a fundamental difference between these approaches. Besides, both approaches set up a formal structure (i.e., a model) between test scores and the attribute being measured. The measurement model developed under common factor theory should be evaluated against observed data for its adequacy by examining the goodness of fit with respect to empirical data before empirical implications of the model can be deduced. Only if the model fit is acceptable then researcher will be allowed to interpret observations as measurements of the latent variables that were hypothesized (Borsboom, 2005). Hence, it should be noted that because the approaches have been developed under different theoretical background, the reliability values obtained from the two methods are not directly comparable.

Six reliability coefficients will be presented in the next section. The first three coefficients represent the general approach for computing reliability of a linear combination when reliabilities of the components are known. The formulas can readily be obtained from the general definitions of reliability and from the properties of the covariance matrix. The last three coefficients represent model based estimates of reliability.

Reliability coefficient based on CTT

This section presents reliability coefficient for a test composed of linear combinations of weighted components under the CTT framework. There are three reliability coefficients presented in this section: stratified alpha, Mosier's coefficient, and Wang and Stanley coefficient. Procedures for calculating the coefficient in this section include: (a) estimating the reliability of each component, (b) calculating the variance of each component, (c) calculating the correlation coefficients among components, and then (d) assigning weights to individual components to form the composite.

Stratified alpha coefficient. The stratified alpha coefficient was introduced by Cronbach, Schoneman, and McKie (1965). This coefficient is suitable to estimating the reliability of measures composed of several subtests, components, facets, or dimensions. The equation for the stratified alpha coefficient is presented here:

$$\alpha_s = 1 - \frac{\sum_{i=1}^{k} \sigma_i^2 \left(1 - r_i\right)}{\sigma_x^2}, \tag{1}$$

where σ_i^2 refers to variance of i component, r_i is reliability of i component, and σ_x^2 is variance of total score (involving all item on the test). This equation indicates that conditions like higher reliability of each component (which means that the test consists of homogeneous items, larger variance of total score) and higher correlation among items or components, will increases possibility to get higher value of stratified alpha. Rae (2007) state that Equation 1 indicates: (a)

Table 1
Variance, reliability and correlations of three dimension of one attribute from fictitious data

Dimension	Number of Items	Variance	Weight	Reliability (alpha)	Correlation		
					A	B	C
A	3	2.20	1	.83		.16	.05
B	3	2.18	2	.84			.09
C	3	2.16	1	.83			
All items		7.87		.71			

when the items within each stratum meet to an essentially tau-equivalent model, the value stratified alpha is equal to the true reliability; (b) if one or more strata have items that meet congeneric model, then stratified alpha will always be a lower bound to reliability; and (c) the greater the variation among the factor loadings, the worse stratified alpha performs as a lower bound.

The following section describes how to compute this coefficient, using the example of a researcher measuring an attitude (comprising three dimensions) toward a political policy. The variance and reliability of each dimension are presented in Table 1. The variance of each dimension is obtained from the total score of items within the same dimension. To obtain the variance of Dimension A, an individual's score for three items within Dimension A (subscale score) is first summed, and the variance of Dimension A is obtained from this score. The variance of the total score is obtained from the variance of the sum of all items on the scale. To obtain this variance, readers should sum all of the items on the scale then compute the variance. In this example, a variance of the total score of 7.87 is obtained from the person's total score on nine items. This value is different when summing three of dimension variances that results in 6.54.

Based on the information presented in Table 1, applying a conventional coefficient alpha to estimate reliability of measurements will underestimate the true reliability (α = .710). In contrast, applying a stratified coefficient alpha will produce a more satisfactory estimation:

$$\alpha = 1 - \frac{2.20 \times (1 - .83) + 2.18 \times (1 - .84) + 2.16 \times (1 - .83)}{7.87} = .861.$$

This finding is consistent with that of Cronbach, Schonemann, and McKie (1965), who used simulation data to demonstrate that the stratified alpha yielded reliability value substantially greater than the alpha itself, since each dimension of the scale measured several independent attributes. Using simulation data, Kamata et al. (2003) found that the stratified coefficient alpha consistently outperformed the traditional coefficient alpha when applied to multidimensional measures. They reported that the stratified alpha coefficient had very low bias in all data conditions of their simulated data. Additionally, the stratified alpha tended to underestimate the true reliability when one component of the scale was miss-specified in the wrong dimension, or when one of the miss-specified components also had a lower reliability than the other components. In general, the stratified alpha estimates the true reliability with low bias, when test components are specified to the correct dimensions.

Rae (2007) explains why the stratified coefficient alpha outperforms the coefficient alpha for multidimensional measures; the stratified alpha can handle an instrument that is fitted to essentially tau-equivalent models, with a possible difference in loading values between dimensions (i.e., congeneric). In contrast, the traditional alpha coefficient tends to perform as a lower-bound estimation of reliability when applied to congeneric measures. If the variation among the factor loadings of dimensions is high, then traditional alpha coefficient will perform worse as a lower-bound estimator.

Mosier's reliability coefficient for composite scores. Mosier (1943) developed a reliability coefficient for measures with multidimensional structures. This coefficient can also be used when measures consist of independent structures reflected in several dimensions. Mosier noted that this coefficient is a general formula for estimating reliability with the possible dimensions weighted. The idea to develop this coefficient arose from an examination of the effect of the interrelationships among the variables on composite validity (Wang & Stanley, 1970). If the dimensions within a measure are mutually uncorrelated, then the reliability of the composite score can be estimated using each dimension's reliability and weighted dimension. To estimate the reliability using Mosier's coefficient, researchers must define the weight, reliability, and variability of each dimension, as well as inter-correlations among dimensions. In this example, we compute the reliability for an aptitude test composed of several subtests, which perform as independent measures or separate dimensions. Unlike the stratified alpha coefficient, the composite score reliability coefficient can accommodate the different weights of each dimension. The equation is:

$$r_{xx'} = 1 - \frac{\sum w_j^2 \sigma_j^2 - \sum w_j^2 \sigma_j^2 r_{jj'}}{\sum w_j^2 \sigma_j^2 + 2 \sum w_j w_k \sigma_j \sigma_k r_{jk}}, \quad (2)$$

where w_j^2 refers to weight for each dimension, r_{jj} is reliability for each dimension, r_{jk} is correlation between each two dimension, and σ_j^2 is variance of each dimension. To compute the composite score reliability, we require information about the reliability, weight value, and variance score of

each dimension, as well as the correlation between dimensions. Using data presented in the Table 1, solving for the numerator and the denominator elements of Equation 2 we obtain:

$$\sum w_j^2\sigma_j^2 = 1^2\times 2.20 + 2^2\times 2.18 + 1^2\times 2.16 = 13.08,$$

$$\sum w_j^2\sigma_j^2 r_{jj'} = 1^2\times 2.20\times .83 + 2^2\times 2.18\times .84 + 1^2\times 2.16\times .83 = 10.94,$$

$$\sum w_j w_k \sigma_j \sigma_k r_{jk} = 1\times 2\times\sqrt{2.20}\times\sqrt{2.18}\times .16 + 1\times 1\times\sqrt{2.20}\times\sqrt{2.16}\times .05 + 2\times 1\times\sqrt{2.18}\times\sqrt{2.16}\times .09 = 1.20.$$

Combining all the above information, we obtain Mosier's reliability coefficient for composite scores equal to .862 which is similar to stratified alpha. The composite score reliability coefficient has several characteristics. First, reliability estimate will achieve 1.00 only if the reliability of each dimension is 1.00. Second, the greater the correlation between the dimensions, the higher the reliability obtained. Third, composite score reliability values are higher than the average reliability of each dimension, with some exceptions. For example, if the reliability of each dimension, variance, and weighted value remain equal and the correlations between the dimensions are close to zero, then Mosier's coefficient will produce a composite reliability equal to the average reliability of the dimensions.

Composite score reliability is defined in terms of the proportion of the total composite variance that serves as an estimation of the true-score variance (Wang & Stanley, 1970). Composite score reliability is an unbiased estimate of the reliability of the general case multidimensional measure for either weighted or unweighted dimensions. One of the advantages of using this coefficient instead of the stratified alpha coefficient is that it accommodates different, appropriate weights for each dimension, which can achieve higher value reliability (Ogasawara, 2009).

There are many alternatives how to weight the components of the measure and add such information to the Mosier's coefficient as well as Wang and Stanley's coefficient explained in the next section. Rudner (2001) proposed two methods for assigning weights to component scores: the implicit approach and the explicit approach. In the implicit method, one can consider adding the raw scores from the components or using item response theory analysis. In the explicit method one can assign weights to individual items of components directly: give more weight to a component that is more difficult (weighting by difficulty), give heavier weights to more reliable components (reliability weighting), or use validity coefficients as weights (validity weighting). The explicit method is relevant for computing Mosier's reliability since researchers can input the weight explicitly.

Several authors have used the composite score of reliability for their multidimensional measures. For example, Harter, Schmidt, Killham, and Asplund (2006) employed this coefficient in a meta-analysis designed to estimate the reliability of the performance measures used in various studies. Since this coefficient is composed of both the reliability of each dimension and the inter-correlations among dimensions, they decided to add updated reliability and inter-correlation among the dimensions to the outcome measures. In this case, they defined the composite performance as an equally weighted sum of customer loyalty, turnover, safety, absenteeism, shrinkage, and financial performance.

Wang and Stanley composite reliability coefficient. Wang and Stanley (1970) stated that when a scale contains a number of component measures (i.e., dimensions), optimal weighting possibly improves the reliability of the composite measure. As a consequence, such a measure will provide a more valid score than if it is merely summed or averaged without weighting. Each component likely has unequal psychometric properties, such as reliability, variance, and intercorrelations with one another. Wang and Stanley assert that since each of the characteristics of the component is reflected in the composite measure, differential weighting for each component would be effective to estimate the reliability:

$$r_{xx'} = \frac{\sum_{i=1}^{n} w_j^2 r_j + \sum_{i=1}^{n}\sum_{j(\neq i)=1}^{n} w_i w_j r_{ij}}{\sum_{i=1}^{n} w_j^2 + \sum_{i=1}^{n}\sum_{j(\neq i)=1}^{2} w_i w_j r_{ij}}, \quad (3)$$

where w_j is weight for j dimension, r_j is reliability for j dimension, and r_{ij} is correlation between i and j dimension. In the case of a measure comprising two dimensions, the reliability of the composite score can be expressed as:

$$r_{xx'} = \frac{w_1^2 r_1 + w_2^2 r_2 + 2w_1 w_2 r_{12}}{w_1^2 + w_2^2 + 2w_1 w_2 r_{12}}. \quad (4)$$

The Equations 3 and 4 indicate that the reliability of the composite score is defined as a function of the weights assigned to the individual dimensions, along with their reliability and correlations with other dimensions. Using data in the Table 1, solving the numerator on right and left side of Equation 3 we obtain:

$$\sum_{i=1}^{n} w_j^2 r_j = 1^2\times .83 + 2^2\times .84 + 1^2\times .83 = 5.02,$$

$$\sum_{i=1}^{n}\sum_{j(\neq i)=1}^{n} w_i w_j r_{ij} = 1\times 2\times .16 + 1\times 1\times .05 + 2\times 1\times .09 = .55,$$

$$\sum_{i=1}^{n} w_j^2 = 1^2 + 2^2 + 1^2 = 6.$$

Applying information given above: 5.02 + .55 / (6 + .55) will produces a Wang and Stanley composite reliability coefficient equal to .850 which approximates stratified alpha as well as Mosier's reliability coefficient for composite scores.

Wang and Stanley composite reliability reaches 1.00 only if every reliability value of each dimension also equals 1.00. Likewise, if it is equal to zero, the reliability value must be zero because the correlation of each dimension is also zero, assuming there is low correlation between two random variables (reliability equal to zero).

Rudner (2001) explained that, in a measure with two dimensions, the lowest possible value of the composite reliability is equal to that of the composite dimension with lowest reliability value. For example, a measure containing two dimensions with reliability values of .7 and .8, respectively, would have the lowest possible composite reliability value of .7. If the two components are correlated, the composite reliability may be higher than the separate reliability of both components.

Reliability coefficient based on common factor model

Reliability coefficients presented in this section are developed under common factor model. The procedure for computing the reliability usually includes: (a) defining measurement model that specifically states the hypothesized relations between the construct being measured and items, (b) making sure that the proposed model fits the data, and (c) inputting related information (e.g., factor loading) into the equation to get the reliability value. Three coefficients are presented in this section: omega coefficient, hierarchical omega coefficient (omega-H), and composite reliability.

Omega coefficient. If a scale is known to contain several independent dimensions, it is possible that the scale score cannot be obtained from a simple summed score, because each dimension has different characteristics. In such a case, each dimension should be treated differently, for example, by weighting dimension measures separately. Depending on how each dimension is weighted to obtain the scale score, the reliability coefficients based on assumptions of unidimensionality and multidimensionality can be distinguished. In some situations, each dimension plays an equal role, and they are thus weighted equally. The coefficient omega accommodates this demand by estimating reliability under equal dimension weighting.

The coefficient omega was proposed by McDonald (1999) using a factor-analytic framework. As with the other reliability coefficients, this coefficient estimates the proportions of a set of indicators that can explain the construct being measured. The author therefore called this coefficient *construct reliability*. However, this term can be confusing, since construct reliability and composite reliability are used interchangeably. For example, Fornell and Larcker (1981) explain that the composite reliability of each latent variable can be used as an estimate of construct reliability. Bacon, Sauer, and Young (1995) called both coefficient alpha and coefficient omega composite reliability or construct reliability coefficients. Composite reliability is used in the context of analysis approach while construct reliability is used in the context of the structure of attribute being measured.

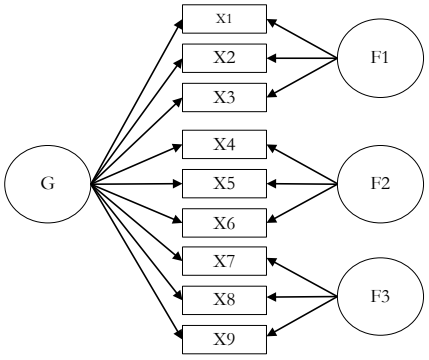

Figure 1. A bifactor model for computing coefficient omega.

The coefficient omega was previously understood as having a unidimensional measurement property, since it is interpreted as an estimate of how much variance in summed scores can be assigned to the single general dimension model of measurement (McDonald, 1999). This model is usually called a common factor model, indicated by a single test composed of multiple dimensions. However, coefficient omega is not a purely unidimensional measurement property (Reise, Moore, & Haviland, 2010); moreover, Heise and Bohrnstedt (1970) suggest coefficient omega as an estimator of reliability for multidimensional measures. The difference between the coefficients alpha and hierarchical omega is the extent to which the reliability estimate is influenced by allowing group factors to figure into true-score variation.

There are two types of coefficient omega: general omega coefficient and weighted or hierarchical omega coefficient. The latter type is calculated by weighting each indicator based on factor loadings. Coefficient omega can be computed using the pattern of coefficients estimated by EFA or CFA (Brunner & Süβ, 2005). Coefficient omega in terms of factor loadings can be expressed as follows:

$$\omega = \frac{\sum_{j=1}^{k}\left(\sum_{i=1}^{p}\lambda_{ij}\right)^2}{\sum_{j=1}^{k}\left(\sum_{i=1}^{p}\lambda_{ij}\right)^2 + \sum_{i=1}^{p}e_i}, \quad (5)$$

where λ_{ij} refers to factor loading of i-indicators on j-factor and e_i refers to unique variance of each indicator. The denominator of the equation is the variance of the summed score of all included items.

Table 2 gives an example using fictitious data comprising nine items from a three-dimensional measure derived from similar data to Table 1. To compute the value of factor loading, CFA is first conducted for a set of scale items. The loading value from CFA, reported by computer programs such as EQS, LISREL, MPLUS, or AMOS, can be used.

Table 2
Standardized factor loading values of nine items from a three-dimensional measure on one general and three specific factors

Dimension	Items	G	F1	F2	F3	Unique
	1	.17	.42			.13
A	2	.17	.42			.10
	3	.15	.42			.14
	1	.20		.38		.12
B	2	.30		.37		.11
	3	.22		.40		.11
	1	.10			.42	.14
C	2	.05			.44	.11
	3	.12			.46	.10

Coefficient omega is model based reliability, therefore reliability values obtained under different models will result in different values as well. Factor loadings presented in Table 2 are obtained from confirmatory factor analysis using bifactor model that estimates general and specific factor simultaneously (see Figure 1). Readers interested in detailed information about bifactor model (usually called *general-specific model* or *nested-factor model*) may consult Reise (2012). Solving for omega from Equation 6 according to information presented in Table 2 yields a value equal to:

$$\sum_{j=1}^{k}\left(\sum_{i=1}^{p}\lambda_{ij}\right)^2 = (.17+.17+.15+.20+.30+.22+.10+.05+.12)^2 + (.42+.42+.42)^2 + (.38+.37+.40)^2 + (.42+.44.+.46)^2$$

$$\sum_{i=1}^{p} e_i = .13+.10+.14+.12+.11+.11+.14+.11+.10 = 1.06,$$

$$\omega = \frac{6.84}{(6.84+1.06)} = .863.$$

Since the proposed model consists of four latent variables (one general and three specific factors), there are four components of summed factor loadings in the numerator of the equation. It should be noted that the denominator of the Equation 5 (6.84 + 1.06 = 7.90) is equal to variance of the summed score of all included items (7.87, see Table 1). Hence, some researchers (i.e., McDonald, 1999) use variance of summed score when computing the value of coefficient omega.

Coefficient omega is widely used by researchers who use SEM with multidimensional constructs (Segars, 1997). Applying this coefficient to parallel or tau-equivalent measurement, which assumes that each item has equal amount of precision of measurement, will obtain an estimate of reliability equal to the coefficient alpha. In contrast, if this coefficient is applied to congeneric measurement, which assumes that each indicator has different amount of measurement precision, the estimated reliability will be higher than the coefficient alpha (Yurdugul, 2006).

McDonald's hierarchical omega coefficient. McDonald (1970) introduced the hierarchical omega coefficient as an estimate of the reliability of measures that consist of several specific unique elements, but for which the general factor still holds. This coefficient has several names, such as *coefficient H* (Hancock & Mueller, 2001), *canonical-factor-regression method coefficient omega* (Allen, 1974), *weighted omega* (Bacon et al., 1995), and *construct reliability* (Brunner & Süβ, 2005). This coefficient modifies the omega coefficient, which is unable to accommodate different weights among dimensions. This coefficient still performs well on unidimensional measures when the congeneric assumption is held, or on multidimensional measures with varied dimensions. The hierarchical omega is expressed as follows:

$$\omega_h = \frac{\left(\sum_{i=1}^{p}\lambda_{ij}\right)^2}{\sum_{j=1}^{k}\left(\sum_{i=1}^{p}\lambda_{ij}\right)^2 + \sum_{i=1}^{p} e_i}, \quad (6)$$

where λ_i is factor loading of *i*-indicators on *j*-factor. The difference between the coefficient omega and omega-H lies in the numerator. The numerator of coefficient omega involves both general and specific components while coefficient omega-H only involves general factors. Again, from Table 2 we can obtain omega-H as follows:

$$\left(\sum_{i=1}^{p}\lambda_{ij}\right)^2 = (.17+.17+.15+.20+.30+.22+.10+.05+.12)^2 = 2.19,$$

$$\sum_{j=1}^{k}\left(\sum_{i=1}^{p}\lambda_{ij}\right)^2 + \sum_{i=1}^{p} e_i = 6.84+1.06 = 7.90,$$

$$\omega = \frac{2.19}{7.90} = .277.$$

The low omega-h indicates that (a) the universe from which one's scale indicators are sampled is multifaceted and (b) the scale scores are a result of the largely independent contributions of what is unique to several of these facets, without much of a contribution from a latent construct that is common to all the facets (Zinbarg, Revelle, Yovel, & Li, 2005).

Coefficient omega and omega-H are different parameters, in most cases omega being always higher than omega-H except in the case of a unidimensional measure. The hierarchical omega coefficient is appropriate for a scale evaluated using SEM, because it allows each component of the measure to be weighted proportional to its true-score variance (Bacon et al., 1995). This coefficient is also appropriate for unidimensional measures, especially when item loadings are not equal. In this case, coefficients omega will yield a greater value than will coefficient alpha when the

number of items is small. Hierarchical omega coefficient can be interpreted as the square of the correlation between the dimensions of the optimal linear composite components; some authors therefore argue that this coefficient produces maximal reliability. However, the reliability value can be improved by appropriately weighting the constituent components. Using simulation data, Bacon and colleagues (1995) found that applying coefficient omega to measures with unidimensional structure, with equal factor loadings for all items, gives the same numerical result as coefficient alpha.

Composite reliability. Raykov and Shrout (2002) provided a coefficient to estimate reliability for composite scores as a generalization of coefficient omega (McDonald, 1999) using SEM approach to the case of nonhomogeneous components. The proposed coefficient can be used to explore the factor structure for a set of items, namely *composite reliability for congeneric measure*. This coefficient is based on SEM and covariance structure analysis methods for scale reliability estimation using congeneric tests (Raykov & Shrout, 2002). The composite reliability coefficient is obtained by taking the ratio between construct and composite variance. The equation is derived from the general concept of intra-class correlation as the ratio of model variance to the total variance, which includes reliability on one hand and the variance explained by the factor(s) on the other hand as two special cases. The term *construct* is used because this coefficient uses CFA. The composite reliability can be calculated using the following equation:

$$r_{xx'} = \frac{\text{var}\left(\sum_{i=1}^{p}\sum_{j=1}^{k}\lambda_{ij}h_i\right)}{\text{var}\left(\sum_{i=1}^{p}\sum_{j=1}^{k}\lambda_{ij}h_i + \sum_{i=1}^{p}E_1\right)}, \quad (7)$$

where λ_{ij} is factor loading of Y_i indicator in factor η_i, η_i is factor i and E_i is error measurement of indicator Y_i. At the moment there is no computer software that directly facilitates computing this coefficient. However, Raykov (1997) created a syntax code for testing models based on SEM using programs such as EQS and LISREL.

To obtain all elements in the Equation 7, a CFA model is set, representing a measure with two dimensions, each consisting of three items (see Figure 2). Three factors are assumed to predict a certain amount of how the construct represents the attribute being measured, with each item also manifesting the composite measure. The composite reliability is obtained by taking the ratio of the construct variance (F4) to the composite variance (F5). Using EQS, we obtained that true composite variance was 1.549 and the construct variance was 1.819. Equation 7 would give a coefficient of the composite reliability of 1.549 / 1.819 = .852.

Raykov (1997) suggested that this coefficient is most trustworthy with large samples because it was developed under a SEM framework. This coefficient should also be applied with care to categorical data with a very limited number of response options. Categorical data are suitably approached by the weighted least squares method of estimation, which performs optimally when applied to large samples; obtaining the composite reliability to estimate parameters on small samples will produce misleading results.

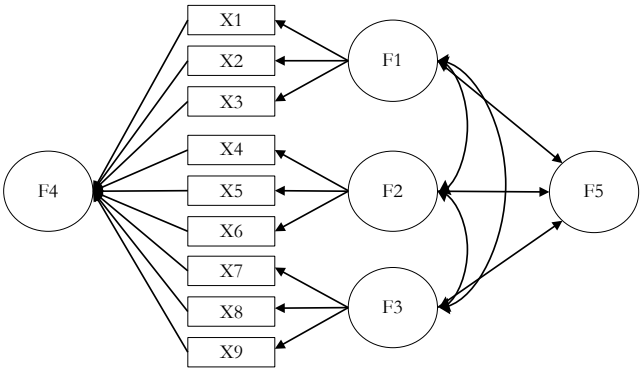

Figure 2. A general structural model for estimating composite reliability.

Using a set of simulated data with $N = 500$ for a scale consisting of six components, Raykov and Shrout (2002) found that the true reliability of the data was .83. Since all model parameters in the datasets were already defined, the true reliability was known. The result obtained by estimating, using the coefficient alpha, was .76, and the composite reliability was .83. These results indicate that the composite reliability can be considered a recommendable reliability estimator, while the coefficient alpha tends to underestimate true reliability.

SUMMARY

As instruments intended to measure attributes, scales usually involve multiple items. Scale items should be interchangeable, since they perform as indicators of the same attribute. Attribute levels can be identified through a composite score that is calculated from the unweighted sum of item scores. Computation of the composite score from a set of scale items is meaningful only if all items hold a unidimensionality assumption (Gerbing & Anderson, 1988). However, in most cases, an inspection of scale dimensionality reveals that items or sets of items have low intercorrelations, even though a core theoretical supposition is that these items assess the same attribute. In such a case, one can infer that the multidimensionality assumption holds. A reliability coefficient based on a unidimensional method (i.e., coefficient alpha) cannot be applied to such a situation. The literature suggests applying an appropriate coefficient in the case of a multidimensional or congeneric measure (e.g., S. Green & Yang, 2009; Raykov, 1997). Recently, these coef-

ficients have gained popularity (Bacon et al., 1995), but few researchers have implemented them in practice.

The reliability coefficients for multidimensional measures described in this article can be divided into two types: reliability coefficient based on CTT approach (i.e., stratified alpha, Mosier's, and Wang-Stanley's coefficients) and common factor model approach (i.e., omega coefficient). As opposed to CTT, common factor model forces researchers to specify the substantive nature of the latent structure underlying items or indicators either in EFA or CFA terms. Reliability coefficient based on CTT approach view reliability as the proportion of true-score variance, without considering the composition of the true score. On the contrary, since common factor approach examines whether covariance shared among components are accounted for by a single or multiple latent factors and whether the latent true score factor loads equally on that factor (Putka & Sackett, 2010), a good reliability estimate requires a fitting model (Green & Yang, 2009) as it is a model-based estimate.

In the case of reliability based on CTT, true score variance for the overall composite score reflects the sum of true score variance for each component of the composite and the sum of covariances between items comprising different components (Putka & Sackett, 2010). In the case of reliability based on common factor model, reliability has been defined as the ratio of the variance due to the common attribute to the total score variance (McDonald, 1999). For higher order factor model, the observed variance of a manifest subtest score is composed of variance attributable to the general (higher) order construct, the variance attributable to the specific constructs, and subtest-specific factors. In this context reliability is mathematically expressed as the proportion of variance in the target construct to observed score variance as McDonald (1999) described. Since there are various variance components decomposed by the model, researchers should decide which variance components contribute to test-score reliability (Bentler, 2006).

Reliability coefficients in this paper can also be distinguished from one another by how weighting is done (explicit weights vs. implicit weights). Two coefficients—Mosier's and Wang-Stanley's reliability coefficient—require researchers to weight dimensions based on scale length, item difficulty, item discrimination, or theoretical bases. The other coefficients (e.g., coefficient omega) are weighted automatically by loading value.

Discussions on unidimensional and multidimensional measurement models still evolve. Multidimensional tests are not only demonstrated by low correlations among test components or items as Osterlind (1998) mentioned that items from unidimensional tests may not correlate highly with each other, but only a single ability accounts for an examinee correctly responding to an item or set of test items. Several authors have warned that the standard (low-dimensional) independent factor CFA model may be much too restrictive for many psychological instruments. Since the empirical research about the influence of model (mis)fit on the factor-analysis based coefficients is almost non-existing, this may pose a limitation to the practical utility of reliability coefficients presented in this paper. On the other hand, coefficient alpha only requires uncorrelated errors as long as it is interpreted properly as a lower bound. Thus, the reliability coefficients presented in this paper can be perceived to serve as an alternative that can be used to enrich the information about the reliability in addition to coefficient alpha.

In summary, we urge researchers to take into account measure dimensionality when calculating reliability. If measures are found to be multidimensional, one of the alternative coefficients described in this article should be applied. The correct coefficient depends on the main method used in the study. For example, studies that use SEM methodology in data analysis should employ reliability coefficients based on CFA. The choice of reliability measure also depends on the design of a study, construct being measured, as well as the measurement model. Bentler (2006) gave a good illustration implemented in our demonstrated data, a single test that consists of three dimensions. A composite made of nine items could be looked at in several ways: a sum across nine items (giving emphasis on general factor) or a sum across three 3-item components (giving emphasis on specific factors). Even the composite score in both approaches is the same; the reliability values computed may differ. Composites made up of more components may or may not have larger reliability coefficients. They will tend to be larger, but do not necessarily need to be, because the assumptions underlying the theory may not be correct (Bentler, 2006).

Nonetheless, this paper is limited to reliability coefficients based on CTT and common factor approach. Recently, with the increasing popularity of item response theory, reliability measures under this approach have caught much attention (Cheng, Yuan, & Liu, 2011). Quite some work has recently been done in the field of multidimensional item-response models (MIRT, see Reckase, 2009). Reliability estimates based on item response theory approach give more emphasis on the error of measurement for each test subject rather than a global index of reliability for the whole test. Hence, several authors feel that devising a single reliability coefficient for a test developed under item response theory approach is inappropriate and misguided. However, when a single coefficient needs to be reported, there are two possibilities (Green, Bock, Humphreys, Linn, & Reckase, 1984). One approach is to define a conditional reliability describing a measurement reliability pertaining to individuals tested with the same precision as other persons with the same level of the measured ability. The other possibility is to define an average or marginal reliability. Marginal reliability is an estimate of the overall reliability of the test based on the average conditional standard errors, which is estimated at different points on the achievement scale, for all examinees.

REFERENCES

Allen, M. P. (1974). Construction of composite measures by the canonical-factor-regression method. In H. L. Costner (Ed.), *Sociological methodology* (pp. 51-78). San Francisco: Jossey-Bass.

Armor, D. J. (1974). Theta reliability for factor scaling. In H. L. Costner (Ed.), *Sociological methodology* (pp. 17-50). San Francisco: Jossey Bass.

Bacon, D. R., Sauer, P. L., & Young, M. (1995). Composite Reliability in Structural Equations Modeling. *Educational and Psychological Measurement, 55*(3), 394-406.

Bentler, P. M. (2006). *EQS 6 structural equations program manual* Encino, CA: Multivariate Software.

Borsboom, D. (2005). *Measuring the mind conceptual issues in contemporary psychometrics* Cambridge: Cambridge University Press.

Bovaird, J. A., & Embretson, S. E. (2008). Modern measurement in the social sciences. In P. Alasuutari, L. Bickman & J. Brannen (Eds.), *The Sage handbook of social research methods* (p. 269). London: Sage.

Brunner, M., & Süβ, H. M. (2005). Analyzing the reliability of multidimensional measures: An example from intelligence research. *Educational and Psychological Measurement, 65*(2), 227-240.

Burt, C. (1938). The analysis of temperament. *British Journal of Medical Psychology, 17*(2), 158-188.

Carmines, E. G., & McIver, J. P. (1981). Analyzing Models with Unobserved Variables: Analysis of Covariance Structures. In G. W. Bohrnstedt & E. F. Borgatta (Eds.), *Social Measurement: Current Issues* (pp. 65-115). Beverly Hills, CA: Sage Publications.

Cheng, Y., Yuan, K.-H., & Liu, C. (2011). Comparison of Reliability Measures Under Factor Analysis and Item Response Theory. *Educational and Psychological Measurement, 72*(1), 52-67.

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York: Harcourt Brace Jovanovich College Publishers.

Cronbach, L. J., Schoneman, P., & McKie, D. (1965). Alpha coefficient for stratified-parallel tests. *Educational & Psychological Measurement, 25*, 291-312.

DeVellis, R. F. (2011). *Scale development: Theory and applications*. Newbury Park: SAGE Publications.

Drolet, A. L., & Morrison, D. G. (2001). Do We Really Need Multiple-Item Measures in Service Research? *Journal of Service Research, 3*(3), 196-204.

Evans, L. D. (1996). Calculating achievement composite scores for regression discrepancy models. *Learning Disability Quarterly, 19*(4), 242-249.

Finn, A. N., Sawyer, C. R., & Behnke, R. R. (2009). A model of anxious arousal for public speaking. *Communication Education, 58*(3), 417-432.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39-50.

Gerbing, D. W., & Anderson, J. C. (1988). An Updated Paradigm for Scale Development Incorporating Unidimensionality and Its Assessment. *Journal of Marketing Research, 25*(2), 186-186.

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*(6), 930-944.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical Guidelines for Assessing Computerized Adaptive Tests. *Journal of Educational Measurement, 21*(4), 347-360.

Green, S., & Yang, Y. (2009). Commentary on Coefficient Alpha: A Cautionary Tale. *Psychometrika, 74*(1), 121-135.

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

Harter, J. K., Schmidt, F. L., Killham, E. A., & Asplund, J. W. (2006). *Q12 Meta-Analysis*. Omaha, NE: Gallup Consulting.

Heise, D. R., & Bohrnstedt, G. W. (1970). Validity, invalidity, and reliability. *Sociological Methodology, 2*, 104–129.

Hendriks, A. A. J., Kuyper, H., Lubbers, M. J., & Van der Werf, M. P. C. (2011). Personality as a moderator of context effects on academic achievement. *Journal of School Psychology, 49*(2), 217-248.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*(2), 109-133.

Kamata, A., Turhan, A., & Darandari, E. (2003, April). *Estimating reliability for multidimensional composite scale ccores*. Paper presented at the American Educational Research Association, Chicago.

Lazarus, R. S. (1991). *Emotion and Adaptation*. Oxford: Oxford University Press.

McCrae, R. R., & Costa, P. T., Jr. (1996). Toward a new generation of personality theories: Theoretical contexts for the five-factor model. In J. S. Wiggins (Ed.), *The five-factor model of personality: Theoretical perspectives* (pp. 51-87). New York: Guilford.

McDonald, R. P. (1970). The theoretical foundations of common factor analysis, principal factor analysis, and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology, 23*, 1-21.

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

Mosier, C. (1943). On the reliability of a weighted composite. *Psychometrika, 8*(3), 161-168.

Ogasawara, H. (2009). On the estimators of model-based and maximal reliability. *Journal of Multivariate Analysis, 100*(6), 1232-1244.

Olson, K., Rogers, W. T., Cui, Y., Cree, M., Baracos, V., Rust, T., . . . Bonville, N. (2011). Development and psychometric testing of the Adaptive Capacity Index, an instrument to measure adaptive capacity in individuals with advanced cancer. *International Journal of Nursing Studies, 48*(8), 986-994.

Osterlind, S. J. (1998). *Constructing test items: Multiple-choice, constructed-response, performance, and other formats*. New York: Kluwer Academic Publishers.

Putka, D. J., & Sackett, P. R. (2010). Reliability and validity. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (pp. 9-49). New York, NY: Routledge.

Rae, G. (2007). A note on using stratified alpha to estimate the composite reliability of a test composed of interrelated nonhomogeneous items. *Psychological Methods, 12*(2), 177-184.

Raykov, T. (1997). Estimation of Composite Reliability for Congeneric Measures. *Applied Psychological Measurement, 21*(2), 173-184.

Raykov, T., & Shrout, P. E. (2002). Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. *Structural Equation Modeling, 9*(2), 195-212.

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. New York: Springer

Reise, S. P. (2012). The Rediscovery of Bifactor Measurement Models. *Multivariate Behavioral Research, 47*(5), 667-696.

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*(6), 544–559.

Rudner, L. M. (2001). Informed Test Component Weighting. *Educational Measurement: Issues and Practice, 20*(1), 16-19.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*(4), 350-353.

Schretlen, D., Benedict, R. H. B., & Bobholz, J. H. (1994). Composite reliability and standard errors of measurement for a seven-subtest short form of the Wechsler Adult Intelligence Scale—Revised. *Psychological Assessment, 6*(3), 188-190.

Segars, A. (1997). Assessing the unidimensionality of measurement: a paradigm and illustration within the context of information systems research. *Omega, 25*(1), 107-121.

Thurstone, L. L. (1931). A multiple factor study of vocational interests. *Personnel Journal, 10*(1931), 198-205.

Wang, M. W., & Stanley, J. C. (1970). Differential Weighting: A Review Of Methods And Empirical Studies. *Review of Educational Research, 40*, 663-705.

Yurdugul, H. (2006). The Comparison of Reliability Coefficients. *Journal of Faculty of Educational Sciences, 39*, 15-37.

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*(1), 123-133.