Journal of Psychoeducational Assessment
29(4) 377–392

DOI: 10.1177/0734282911406668
http://jpa.sagepub.com
SAGE

Coefficient Alpha: A Reliability Coefficient for the 21st Century?

Yanyun Yang[1] and Samuel B. Green[2]

Abstract

Coefficient alpha is almost universally applied to assess reliability of scales in psychology. We argue that researchers should consider alternatives to coefficient alpha. Our preference is for structural equation modeling (SEM) estimates of reliability because they are informative and allow for an empirical evaluation of the assumptions underlying them. An example is presented to illustrate the advantages of SEM estimates of reliability.

Keywords

coefficient alpha, reliability, structural equation modeling

Coefficient alpha is the most popular coefficient reported to support the reliability of a scale (e.g., Cronbach, 2004). We consider four reasons why researchers have so frequently chosen to report coefficient alpha:

1. Coefficient alpha is easy to interpret. To the degree coefficient alpha approaches 1.0, a scale demonstrates greater internal consistency. Although coefficient alpha may be an underestimate of reliability if assumptions are violated, it is simple and even desirable to include a caveat about alpha being an underestimate in that readers then understand that the true reliability might be greater.
2. An advantage of coefficient alpha over many other reliability estimates is that it is objective and does not require subjective decisions; therefore, it is straightforward to use. For example, to compute a split-half coefficient, researchers must decide how to split a scale into two halves. Coefficient alpha does not require the splitting of scales. Moreover, to compute a test-retest coefficient, researchers must decide how long a time should occur between the test and the retest. Coefficient alpha does not require researchers to administer a scale on two occasions.
3. Coefficient alpha is useful in making revisions to scales. Based on coefficient alphas that would be obtained if items are deleted, researchers can make decisions about what

[1]Florida State University, Tallahassee
[2]Arizona State University, Tempe

Corresponding Author:
Yanyun Yang, Educational Psychology and Learning Systems, Box 306-4453, Florida State University, Tallahassee, FL 32306-4453
Email: yyang3@fsu.edu

items should be included or excluded from a scale. Also, coefficient alpha is useful in assessing the potential validity of a scale (i.e., the square root of alpha).

4. Thousands of psychological researchers who report coefficient alphas or review manuscripts that include coefficient alphas cannot be wrong. Because of its popularity, researchers have developed a normative framework for interpreting the values for coefficient alphas as small, medium, and large.

We will demonstrate in this article that these reasons lack merit and are based on a faulty understanding of coefficient alpha. Rather than routinely choosing coefficient alpha as a reliability estimate, we argue that researchers should carefully choose among reliability coefficients. More specifically, we discuss why researchers should conduct preliminary analyses, such as structural equation modeling (SEM), to understand the internal structure of a scale before choosing an internal consistency estimate of reliability. Preliminary analyses may or may not lead researchers to choose coefficient alpha. For example, if SEM methods suggest that a measure is multidimensional, then coefficient alpha would be inappropriate, and SEM reliability estimates could be computed that are perfectly consistent with the preliminary analyses. We try to present our viewpoint succinctly and nontechnically. Due to space limitations, we must assume that readers have a basic understanding of reliability and validity. For a more technical discussion of coefficient alpha and alternative reliability coefficients, readers should read a recent set of articles in *Psychometrika* (Bentler, 2009; Green & Yang, 2009a, 2009b; Revelle & Zinbarg, 2009; Sijtsma, 2009).

Reliability and Coefficient Alpha

In this section, we begin by presenting definitions for reliability. Next, we introduce coefficient alpha as a reliability coefficient, assumptions underlying coefficient alpha, and consequences of violating assumptions. We end this section by considering interpretation of coefficient alpha in practice.

Reliability

An observed score on a measure for an individual (X) can be conceptualized as being a sum of two unobservable scores: a true score (T) and an error score (E); that is, $X = T + E$. The true and error scores represent, respectively, the replicable and inconsistent parts of an observed score and are assumed to be uncorrelated with each other. *Reliability* is defined in the population as the ratio of true score variance to observed score variance, $\rho_{XX'} = \sigma_T^2 / \sigma_X^2$, or, alternatively, in terms of error score variance as $\rho_{XX'} = 1 - \sigma_E^2 / \sigma_X^2$. In theory, we can also think of reliability as the correlation between a scale administered on two occasions, with zero time between administrations (so true scores for individuals do not change); this view of reliability underlies the choice of the symbol for reliability as $\rho_{XX'}$.

Alpha and Its Assumptions

In psychology, we frequently are interested in the reliability of scale scores that are computed by summing item scores. To assess reliability based on a single administration of a scale, researchers can compute coefficient alpha. Coefficient alpha in the population is $\alpha = (n^2\bar{\sigma}_{ii'})/\sigma_X^2$, where n is the number of item, $\bar{\sigma}_{ii'}$ is the mean covariance between pairs of items, and σ_X^2 is the variance of observed scale scores. For sample data, an estimate of the population coefficient alpha can be computed by substituting sample values for the mean covariance between pairs of items and the variance of observed scale scores: $\hat{\alpha} = (n^2\bar{C}_{ii'})/S_X^2$.

If item scores meet certain assumptions, coefficient alpha is equal to reliability in the population,[1] although it may be higher or lower than the population reliability for any particular sample. There are three primary assumptions underlying the derivation of coefficient alpha as equal to reliability in the population (Novick & Lewis, 1967):

1. *Classical item-score assumption:* Item scores are a sum of item true and error scores.
2. *Tau equivalency assumption*[2]*:* The same true scores underlie all items and equally contribute to all item scores. Within the framework of factor analysis, items have equal loadings on a single underlying factor.
3. *Uncorrelated-errors assumption:* Item error scores between any pair of items are uncorrelated.

Violations of Assumptions and Consequences

All three of these assumptions are likely to be violated to some degree in practice, and, therefore, the accuracy of coefficient alpha as an estimate of reliability is problematic.

Classical item-score assumption. The classical item-score assumption is in conflict with the view held by most psychometricians about item scores. Scores on items are typically limited to a small number of values, for example, 0 or 1 for right–wrong items or 1 through 5 for agree–disagree Likert-type items. For items with limited response scales, it is difficult to conceptualize how item scores are a simple sum of item true and error scores. For example, with right–wrong items, a modern view (i.e., item response theory) is that the probability of obtaining a correct answer on an item is curvilinearly related ("s-shaped" curve) to true scores of individuals. See Embretson and Reise (2000) for further discussion of this topic. Although it is unclear whether violation of this assumption has an impact on the accuracy of coefficient alpha, we will see that it does affect the way that we assess violation of the other two assumptions.

Tau equivalency assumption. In an absolute sense, items on a scale are unlikely to be tau equivalent for two reasons. First, if a scale is truly unidimensional, it is doubtful whether a single underlying factor would contribute exactly the same amount to every item (Cortina, 1993). Second, scales are not likely to be unidimensional. In fact, Reise and his colleagues (Reise, Morizot, & Hays, 2007; Reise, Waller, & Comrey, 2000) argued that noncognitive scales are unidimensional only if they assess narrow constructs or, more provocatively stated, trivial constructs. It is preferable if a bifactor model underlies items on these scales such that all items are strongly associated with a general factor and subsets of items are associated with group factors. In other words, items should be predominately a function of a general factor (i.e., essentially unidimensional).

If the tau equivalency assumption is violated (and uncorrelated errors is assumed), then coefficient alpha is less than reliability in the population and will tend to be too small when computed in samples. When tests are unidimensional, the underestimation of reliability by alpha is minimal, unless one or two loadings are considerably greater than the others (Green & Yang, 2009a; Raykov, 1997b). However, the degree of bias under these conditions can be substantial (e.g., 10% underestimate) for scales with small numbers of items (e.g., six-item scale). When tests are multidimensional, the degree of bias may range from minimal to considerable. For example, Green and Yang (2009a) demonstrated that the underestimation may be as high as 19% for a six-item test with a relatively weak general factor and strong group factors. The degree of bias appears to be more substantial for scales with fewer items (Feldt & Qualls, 1996; Green & Yang, 2009a). From a practical viewpoint, it seems likely that items on well-developed scales are more likely to be approximately tau equivalent and to demonstrate less bias (Green & Yang, 2009a; Raykov, 1997b).

Uncorrelated errors assumption. Although this assumption is ignored by many researchers who discuss coefficient alpha, some psychometricians perceive this assumption as very problematic.

Guttman (1945) warned many years ago about difficulties with meeting the uncorrelated errors assumption for internal consistency reliability estimates in general. Rozeboom (1966) was very explicit in his admonitions: ". . . however pleasant a mathematical pastime it may be to shuffle through the internal statistics of a compound test in search of a formula which gives the closest estimate of the test's reliability under conditions of uncorrelated errors, this is for practical applications like putting on a clean shirt to rassle a hog" (p. 415). More recently, Cronbach (2004) stated, "One can rarely assert, then, that violations of independence are absent, and it is burdensome (if not impossible) to assess the degree and effect of nonindependence" (p. 402).

There are potentially many ways that this assumption may be violated. Errors may be correlated if a test with right–wrong items is speeded (Cronbach, 2004; Rozeboom, 1966) or if subsets of items on a scale are associated with different stimuli, such as items assessing reading comprehension for different paragraphs (e.g., Steinberg, 2001; Steinberg & Thissen, 1996; Wainer & Kiely, 1987; Yen, 1993). Potentially, the ordering of items could introduce correlated errors, particularly if similarly worded items were adjacent to each other (Cronbach, 2004; Green & Hershberger, 2000). In addition, transient effects associated with taking a set of items on one occasion may introduce correlated errors (Becker, 2000; Green, 2003). In most cases, the errors are likely to be positively correlated, and the effect is inflated coefficient alphas. Potentially, the effect could be quite profound (e.g., Fleishman & Benson, 1987; Komaroff, 1997; Maxwell, 1968; Miller, 1995; Raykov, 1998; Zimmerman, Zumbo, & Lalonde, 1993), but the size of the effect is less clear in practice because few researchers have routinely investigated scales for correlated errors. See Green and Hershberger (2000) and Green and Yang (2009a) for further discussion of this topic.

Based on our presentation of violations of assumptions, we can now respond to the following item: Coefficient alpha may be (a) negatively biased, (b) relatively unbiased, (c) positively biased, or (d) all of the above. The answer is "all of the above." For any one scale, the degree that it is negatively or positively biased is dependent on the degree that the tau equivalency and uncorrelated errors assumptions are violated in combination.

Interpretation of Coefficient Alpha

As we have discussed, coefficient alpha is a statistic that assesses the reliability of a scale based on its internal consistency. Unfortunately, many researchers fail to include the word "reliability" when reporting coefficient alpha and simply state that it is a statistic assessing internal consistency (Green & Thompson, 2003). Referring to coefficient alpha as an internal consistency index is problematic for a number of reasons. One difficulty is that some researchers use the terms internal consistency and homogeneity interchangeably (Cortina, 1993). We adopt the distinction made by Cortina (1993), Green, Lissitz, and Mulaik (1977), and Schmitt (1996) in discussing coefficient alpha that internal consistency is the interrelatedness among items (i.e., the average interitem correlation), whereas homogeneity is the degree that items are unidimensional. Coefficient alpha does not reflect the degree that a scale is homogeneous (e.g., Cortina, 1993; Green et al., 1977; Miller, 1995; Schmitt, 1996). A scale can be unidimensional and have a low or a high coefficient alpha; a scale can be multidimensional and have a low or a high coefficient alpha. In addition, although coefficient alpha is sensitive to the internal consistency of a scale, it is heavily influenced by the number of items on it. To illustrate, let us say all items have variances of 1, and correlations between all items are uniformly .3. Although this set of items has the same degree of internal consistency (i.e., average interitem correlation of .3), coefficient alpha would be .46 for a two-item scale and .82 for a five-item scale. Clearly, coefficient alpha should not be reported as an index of internal consistency.

Multiple researchers have made suggestions about acceptable levels of coefficient alpha as a reliability coefficient. Nunnally (1978) is frequently quoted for the following cutoff values: .70 for scales in the initial level of development, .80 for basic research scales, and .90 as the minimal level for scales used for clinical purposes and .95 as an ideal level for these scales. Others (e.g., George & Mallery, 2003; Kaplan & Saccuzzo, 2009) report similar cutoff values. Interestingly, these cutoffs appear to be consistent with normative findings concerning coefficient alpha. Peterson (1994) examined studies that report coefficient alpha primarily in the business and psychology literatures. Across 4,286 alpha coefficients, the mean coefficient alpha was .77. Seventy-five percent of the alphas were .70 or greater, whereas only 14% reached or exceeded .90. Some researchers report low coefficient alphas (e.g., .65) as acceptable and justify this conclusion by indicating that their scales contain few items (Schmitt, 1996). It is true that alpha is directly affected by scale length. Nevertheless, we cannot make accurate judgments about respondents based on scales with low reliabilities, regardless of their length. Finally, judgments about cutoffs for coefficient alpha become less meaningful to the extent that computed alphas are inaccurate estimates of reliability due to violations of assumptions.

Coefficient alpha is sometimes used to make decisions about what items to include or exclude from a measure. Reliability analysis in SPSS and PROC CORR in SAS report coefficient alpha for a scale if an item is deleted. Presumably, researchers delete items from scales if alphas increase in magnitude or even remain relatively unchanged. We have a number of concerns about using this approach. First, researchers might incorrectly apply this method with the purpose of developing unidimensional measures. As we have discussed, coefficient alpha is an inappropriate statistic for assessing unidimensionality. Second, researchers might use this method to create maximally reliable measures. However, if the assumptions of coefficient alpha are violated, the wrong items could be deleted or retained in that coefficient alpha might be a poor estimate of reliability. For example, items might be retained on a scale if they are similarly worded and are located near each other on the scale. It is reasonable to hypothesize that these items are highly correlated not because they, in combination, create a reliable scale but rather the errors for these items are correlated. Third, under some conditions, applying this method could result in a scale assessing a very narrow construct and, thus, possessing unsatisfactory validity (Reise, et al., 2000). In other words, the effect might be a trade-off: an increase in fidelity, but at the expense of a loss in bandwidth (Cronbach & Gleser, 1965).

Alternatives to Coefficient Alpha

A variety of methods are available for computing reliability.[3] We will discuss briefly alternative traditional methods for computing reliability. We will then consider estimates using SEM. These methods can be seen as an extension of approaches used to judge whether assumptions underlying coefficient alpha (as well as other reliability coefficients) are met.

Standard Alternatives: Test-Retest, Equivalent Forms, and Split-Half Coefficients

One alternative to coefficient alpha is a test-retest coefficient. It allows for the assessment of transient errors (e.g., amount of noise during administration of a scale or amount of sleep prior to taking a test) that have differential effects for the two administrations of a test (Becker, 2000; Hunter & Schmidt, 1996). The dilemma faced by researchers using a test-retest coefficient is choosing a length of time between administrations of a scale that allows for differential transient errors but does not permit changes in true scores. This judgment about length of time between the

test and the retest is further complicated by the potential, undesirable effect of test takers' memory of item responses at the first administration on their item responses at the second administration. As discussed in most introductory measurement books, researchers must choose a length of time that is sufficiently long to minimize memory effects but that is not too long to minimize changes in the construct(s) being assessed. As test-retest reliability assumes stability of the measured latent constructs over time, this method is not appropriate for constructs that tend to change over time (e.g., mood).

The use of an equivalent-forms coefficient allows for the assessment of transient errors but avoids problems associated with memory of item responses. However, in most applications, an equivalent form is unavailable and, thus, this coefficient is not typically perceived as an option. Recently researchers have suggested some new approaches for computing coefficients that assess transient errors and minimize memory effects which are adaptations of equivalent-forms and test-retest coefficients (e.g., Becker, 2000; Green, 2003).

Another traditional alternative is the split-half coefficient. This internal consistency reliability coefficient is computed based on results from a single administration of a scale. Accordingly, split-half coefficients do not assess transient errors (see Becker [2000] for an interesting adaptation of a split-half coefficient that assesses transient errors). To calculate a split-half coefficient, a scale is divided into two halves that are as equivalent as possible. At this point, two methods can be used to compute the split-half coefficient:

Method 1: A correlation is computed between the two half-scale scores and then the Spearman–Brown prophesy formula is applied to estimate a scale's reliability. This approach requires the two halves to be parallel.

Method 2: Coefficient alpha can be computed, treating the half-scale scores as items: $\hat{\alpha}_h = (n^2 C_{hh'})/S_X^2$ where h and h' are two halves of a scale and $C_{hh'}$ is the covariance between the two halves. For this method, the halves must be tau equivalent. The split-half coefficient was adapted by Raju (1970) to allow for splitting scales with an odd number of items, such that any split would have one more item in one split versus the other.

Researchers who wish to compute a split-half coefficient are faced with the decision about how to split a scale into equivalent halves (Osburn, 2000). Researchers may choose to avoid this decision and calculate coefficient alpha, which is the mean of all possible split-half coefficients computed using the second method (Cronbach, 1951). However, by abrogating the responsibility for making this decision, they are choosing a less than optimal estimate of reliability if items are not tau equivalent. In fact, if items are not tau equivalent, then all split-half coefficients are less than or equal to a scale's reliability (assuming the correlated errors assumptions was not violated). (See Thompson, Green, & Yang [2010] for illustrations.) Based on this rationale, Callendar and Osburn recommended splitting a scale in all possible ways and choosing the largest split-half coefficient as the reliability estimate (i.e., maximal split-half coefficient; Callendar & Osburn, 1977a, 1977b, 1979; Osburn, 2000). Unfortunately, although this method is appropriate at the population level, it can yield positively biased reliability estimates in practice due to capitalization on the chance characteristics of a sample. The bias is corrected if the maximal split-half coefficient is cross-validated, but cross-validated coefficients tend to be unstable across samples (Thompson et al., 2010).

SEM Methods for Computing Reliability

Besides split-half coefficients, we restrict our focus to internal consistency methods for estimating reliability using SEM. It is important to note that reliability could be estimated based on the

results of exploratory factor analysis (Revelle & Zinbarg, 2009), although we find SEM methods more flexible.

We use a two-step approach in estimating reliability using SEM. In Step 1, a judgment is made about which confirmatory factor analytic (CFA) model (i.e., the measurement model in SEM) best represents the data. In Step 2, reliability is estimated based on the model found in Step 1. We next discuss these steps in greater detail.

Step 1: Assessment of CFA models. In this step, a series of CFA models are evaluated to determine the most appropriate model based on a researcher's understanding of the scale of interest and the empirical fit of the models to the item data. We make decisions among models using global fit indices (e.g., chi-square, Comparative Fit Index [CFI], and root mean square error of approximation [RMSEA]) and changes in these fit indices as well as local fit indices (e.g., interpretability of the factor loadings). In most applications, a summed score is not meaningful unless a relatively strong general factor underlies items on a scale; that is, a scale must be essentially unidimensional to be interpretable. Accordingly, we restrict our focus to models that include a general factor, although these models may also include group factors. Moreover, for a scale to be essentially unidimensional, the loadings on the general factor must be relatively high. We can proceed to Step 2 if the results of Step 1 suggest that items are essentially unidimensional.

In conducting Step 1, researchers may start with a single-factor model that allows for unequal factor loadings (i.e., a congeneric model). If this model fits relatively well, they should evaluate the fit of a more constrained model that requires all factor loadings to be equal, that is, a tau-equivalent model. If the fit of the congeneric and the tau-equivalent models are approximately the same, researchers should choose the more parsimonious model, the tau-equivalent model; otherwise, researchers should choose the congeneric model. If the congeneric model fails to fit, bifactor models (i.e., inclusion of a general factor associated with all items and one or more group factors associated with subsets of items) should be assessed. The choice of group factors within the bifactor model should be primarily guided by the researchers' understanding of the items and the respondents in the sample. In specifying a bifactor model, the general and group factor(s) should be constrained to be uncorrelated to minimize problems with estimation and difficulties with interpretation. Researchers should select a bifactor model if it is conceptually meaningful and fits adequately. If no bifactor models meet these criteria, researchers should consider revisions to the scale to achieve essential unidimensionality.

For any of the described models, covariances between errors (i.e., allowing for correlated errors) could be included. These covariances between errors should be included only if they can be conceptualized as due to measurement error, for example, if different subsets of items are associated with different visual stimuli (e.g., maps). If covariances between errors are due to true score variability, then these covariances should be replaced with one or more group factors as part of a bifactor model. The variances of these group factors would then be as defined as part of true score variance. See Green and Hershberger (2000) for a more in-depth discussion on differentiating between true and error score variances as the source of covariances between errors.

Prior to fitting these various models, a choice must be made as to whether to use a linear or nonlinear SEM approach. In practice, item scores are almost always ordered categorical, with a limited number of response points (e.g., 0 or 1 depending on whether the item is incorrect or correct on an achievement test). With the standard linear approach, CFAs are conducted on a covariance matrix among the items, typically applying maximum likelihood methods. This approach can yield incorrect fit indices, biased factor loadings, and spurious factors, although linear factor analyses can yield more accurate results as the number of scale points increases (Green, 1983; Green, Akey, Fleming, Hershberger, & Marquis, 1997; Muthén & Kaplan, 1992). To obviate these difficulties, nonlinear SEM methods should be used with ordered categorical data (e.g., Finney & DiStefano, 2006; Flora & Curran, 2004; Muthén & Muthén, 1998-2009). These methods require

substituting a matrix of polychoric correlations among items for a matrix of item covariances as the input for CFA. Polychoric correlations estimate the correlations between the normally distributed continuous item scores hypothesized to underlie the observed ordered categorical item scores. (Tetrachoric correlations are a special type of polychoric correlations for items with binary response scales.) The CFA model then is estimated using weighted least square with mean and variance adjustment (WLSMV; Muthén & Muthén, 1998-2009) rather than maximum likelihood (Lei, 2009). Flora and Curran (2004) and Lei (2009) demonstrated with computer-generated data that WLSMV yields relatively accurate results when the model is correctly specified, although slightly biased test statistics and standard errors are obtained when large models are estimated with small samples.

Step 2: Estimation of the reliability coefficient. In Step 2, a reliability coefficient is computed based on the results of Step 1. Researchers may take one of two routes in this step. With the first route, they could choose among traditional reliability coefficients using the results of Step 1. For example, coefficient alpha may be computed if the findings from Step 1 suggest a tau-equivalent model with no correlated errors (i.e., without covariances among errors). Alternatively, if the CFA suggested a unidimensional model but with diverse factor loadings, they might compute a split-half reliability coefficient, splitting items into two approximately equivalent halves based on the CFAs performed in Step 1 (Green & Yang, 2005).

With the second route, researchers may compute an SEM reliability coefficient using the parameter estimates from the model chosen in Step 1. An SEM coefficient may be computed based on a linear CFA model (e.g., Bentler, 2009; Green & Yang, 2009a; McDonald, 1978, 1999; Raykov & Shrout, 2002) or on a nonlinear CFA model (Green & Yang, 2009b). As described above, a nonlinear model is more appropriate if items are ordered categorical, particularly if the number of response categories is limited.

With a linear model, the reliability coefficient is a ratio of a scale's estimated true score variance to the scale's estimated total score variance. These estimated variances are a function of the fitted linear model. The scale's estimated true score variance is due to the factor(s), whereas the scale's estimated total score variance is due to the factors and the errors. This reliability estimate was referred to as *coefficient omega* (ω) by McDonald (1978) and renamed as *omega total* (ω_T) by Revelle and Zinbarg (2009). EQS, by default, reports a linear SEM reliability coefficient (labeled *reliability coefficient rho*). In other SEM software programs, the estimated variances in the numerator and denominator of the linear SEM reliability coefficient can be computed by the programs and do not need to be computed by hand. In the syntax for the SEM programs, two "phantom factors" are defined to represent the scale true score and the scale total score (Raykov, 1997a; Raykov & Shrout, 2002). The phantom factor for scale true score is delineated as follows: (a) sum loadings for each factor, (b) multiply summed loadings for a factor by that factor, and (c) sum products from results obtained in (b). The phantom factor for scale total score is defined simply as the sum of the observed items. To compute the linear SEM estimate, the model-based variances for these two phantom factors should be obtained from the output of the programs. Example of syntax for these SEM programs is available at http://myweb.fsu.edu/yyang3/yang.html. As demonstrated by Yang and Green (2010a), the linear SEM estimates yield biased estimate if the structural model is misspecified. In addition, overspecification of a model is likely to yield overestimates of reliability, although the bias is minimized with large sample sizes.

A nonlinear SEM estimate of reliability is hypothetically the correlation between a scale and its parallel form with zero time between administrations, although the data are obtained from a scale administered on a single occasion. The formula for computing this coefficient is complex, but an SAS program is available to calculate it (Green & Yang, 2009b). Green and Yang (2009b) demonstrated at the population level under a variety of conditions that this coefficient performs equally well or better than the linear coefficient or coefficient alpha. A recent simulation study showed that the nonlinear SEM reliability coefficients perform reasonably well and better than linear

SEM reliability coefficients at both population and sample levels when the underlying distributions of the item scores are normal or deviate moderately from normality (Yang & Green, 2010b). Further research is necessary to understand how it performs if the model is misspecified.

Clearly, a difficulty with SEM estimates of reliability (linear or nonlinear) is that they can yield inaccurate results if the model on which it is based is incorrect. However, with these estimates, data guide the decision about choice of a model. In contrast, researchers who report coefficient alpha typically assume a tau-equivalent model without any empirical justification. Three additional interrelated points are worth noting. First, a reliability coefficient based on a CFA has richer interpretational value than a traditional reliability coefficient in that the results of the CFA clarify our understanding of the coefficient and its limitations. Second, one aspect of our understanding brought about by the CFA is that reliability evaluates the contribution of not only the focal factor (i.e., the general factor) but also additional factors (i.e., group factors) that must be considered in the interpretation of a scale. Third, the computation of the SEM coefficient requires a judgment concerning the validity of a scale in the sense that a scale should be essentially unidimensional before computing it.

Empirical Example

We next present an empirical example using TIMSS data (collected in 2007) to illustrate coefficient alpha, split-half coefficients, and SEM reliability coefficients. Data from U.S. fourth graders consisted of 7,831 cases. For our illustration, we selected a random sample of 293 cases, with no missing values. A seven-item scale was used to measure fourth graders' attitude toward learning mathematics in school and, more specifically, positive affect toward learning mathematics (PAM; Items 1-3) and self-confidence in learning mathematics (SCM; Items 4-7). Each of the seven items was rated on a 4-point scale (1 = *agree a lot to* 4 = *disagree a lot*). Items 1, 3, 4, and 7 were worded positively, but response scales were reversed for our example such that higher scores on all items indicated positive attitude toward learning mathematics. Table 1 presents the items and item statistics, including means, standard deviations, Pearson correlations, and polychoric correlations for this random sample. The means of the interitem Pearson correlations and polychoric correlations were .394 and .479, respectively. The raw data for this sample is available at http://myweb.fsu.edu/yyang3/yang.html.

Alpha and Split-Half Coefficients

Coefficient alpha would typically be computed for this type of data. Its value was .815. According to Nunnally (1978), the scale has adequate reliability for basic research. No further interpretation of coefficient alpha would be provided by most applied researchers.

Alternatively, a split-half coefficient could be reported. Given seven items, we considered splits with three items in one half and four items in the other half (Raju, 1970). For all possible 35 splits, the coefficients ranged from .658 to .885, with a mean of .815. As expected, the mean split-half coefficient was equal to coefficient alpha. From these results, it is clear the choice in splitting scales can lead to very different results and interpretations.

We suspect that none of the splits yielded equivalent halves in the population. Thus, assuming no correlated errors, all 35 coefficients could be viewed as lower bound estimates of reliability. Within this framework, the coefficient of .885 might be viewed as the "best" lower bound estimate (based on Items 1, 2, 4, and 5 in one half and Items 3, 6, and 7 in the other). It is important to cross-validate the result to avoid capitalizing on the chance characteristics of the sample. To achieve this goal, we randomly selected a second sample with 289 participants and computed for this sample a split-half coefficient of .865 using the same split of items. As expected, this coefficient was somewhat lower than the largest split-half coefficient based on the original sample, but larger than coefficient alpha.

Table 1. Means, Standard Deviations, Pearson Correlations (Lower Triangle), and Polychoric Correlations (Upper Triangle) for the Seven Items Addressing Attitude Toward Learning Mathematics in School

	Items						
	1	2	3	4	5	6	7
1. I enjoy learning mathematics[a]		.679	.860	.421	.297	.471	.410
2. Mathematics is boring	.559		.645	.317	.323	.485	.212
3. I like mathematics[a]	.778	.520		.482	.330	.422	.426
4. I usually do well in mathematics[a]	.351	.238	.401		.516	.684	.565
5. Mathematics is harder for me than for many of my classmates	.254	.272	.275	.409		.662	.403
6. I'm just not good at mathematics	.376	.390	.330	.538	.547		.455
7. I learn things quickly in mathematics[a]	.358	.171	.375	.464	.324	.351	
M	3.119	3.130	3.106	3.362	2.911	3.369	3.034
SD	0.995	1.081	1.059	0.749	1.113	0.926	0.921

a. Items that were positively worded but reverse scaled.

Linear SEM Methods for Computing Reliability

We demonstrate linear SEM methods for computing reliability following the two-step approach. With these methods, models were fit to the covariance matrix with maximum likelihood methods assuming a linear relationship between factors and items.

Step 1: Assessment of CFA models. The seven-item scale was intended to measure fourth graders' attitude toward learning mathematics in school. We began by fitting a one-factor model (Model 1), as depicted in Figure 1a. The variance of the factor was arbitrarily fixed at 1 for this model as well as all other CFA models presented in the article. This model fit the data poorly, $\chi^2(14) = 201.18$, $p < .01$, CFI = .76, RMSEA = .21, and standardized root mean square residual [SRMR] = .11. As the one-factor model failed to fit the data adequately, we did not examine the more constrained one-factor model with equal factor loadings (i.e., a tau-equivalent model).

Next, we considered a series of bifactor models.[4] These models allow for the scale to be essentially unidimensional by not only including a general factor but also specifying group factors. The first bifactor model (Model 2) included a group factor for Items 1 to 3 as well as a general factor, as shown in Figure 1b. The group factor represents PAM. This model fit reasonably well, $\chi^2(11) = 48.07$, $p < .01$, CFI = .95, RMSEA = .11, and SRMR = .04. The loadings on the general factor ranged from .45 to .71, whereas the loadings on the group factor were .77, .49, and .72.

In that Items 2, 5, and 6 were negatively worded, we chose to include a second group factor in the bifactor model, as shown in Figure 1c (Model 3). This second group factor was also suggested by the modification indices based on Model 2. Model 3 fit well, $\chi^2(8) = 14.83$, $p = .063$, CFI = .991, RMSEA = .05, and SRMR = .02. Compared with the previous bifactor model, the CFI increased, whereas both the RMSEA and SRMR decreased considerably. The chi-square difference test also showed that Model 3 fit significantly better than Model 2. The loadings on the general factor ranged from .33 to .62; the loadings on the PAM group factor were .73, .58, and .72; and the loadings on the negatively worded items were .36, .40, and .51. It should be noted that the same fit would be achieved by including covariances among errors for Items 2, 5, and 6 (Model 4, as depicted in Figure 1d), as found with Model 3. However, we prefer the specification of a second group factor (i.e., Model 3) as opposed to covariances among errors in that

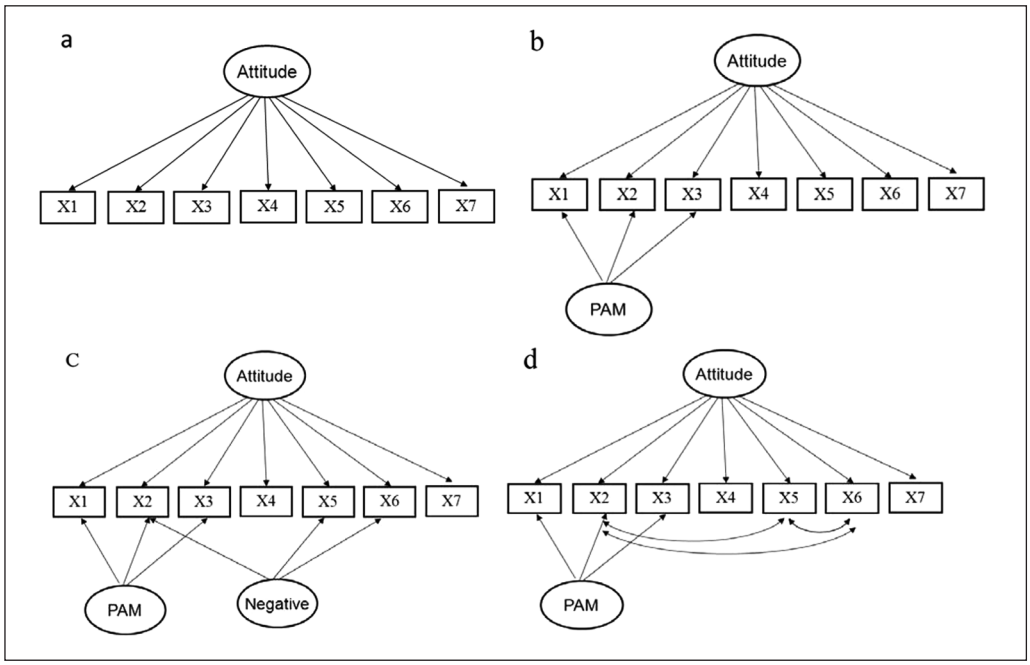

Figure 1. Four possible confirmatory factor analysis models for the example data

the factor can be interpreted based on the content of the items. In other words, we may view it as an irrelevant source of variance and detracting from the scale's validity, but not due to measurement error.

Step 2: Estimation of the reliability coefficient. Models 2 and 3 demonstrated adequate fit (as summarized in Table 2) and have justification substantively. The percentages of estimated common variance (i.e., variance due to all factors) accounted for by the general factor were 79.6 and 70.8 for Models 2 and 3, respectively, indicating a moderately strong contribution of the general factor. The linear SEM coefficients for Models 2 and 3 were .860 and .876, respectively. It is worth noting that the linear SEM coefficient for Model 4 was .805, substantially less than the other estimates in that correlated errors contribute to error score variance, not true score variance. Mplus, SAS, and EQS syntax for computing true score variance, total score variance, and reliability coefficient for Models 2 through 4 can be found at http://myweb.fsu.edu/yyang3/yang.html.

We would argue that the reliability coefficient based on Model 3 is preferable to the reliability estimated based on the other models in that all three factors are interpretable based on the content of the items. From our results, we can see the model specification has a demonstrable effect on the size of the reliability coefficients.

Nonlinear SEM Methods for Computing Reliability

In comparison with nonlinear SEM methods, linear SEM methods are better known and understood and, thus, are more likely to be applied in practice. In that the response scale for the seven items is limited to four categories, the nonlinear method would appear to be preferable. We demonstrate next nonlinear SEM methods for computing reliability following the two-step approach. With these methods, models were fit to a matrix of polychoric correlations among items using WLSMV in Mplus 5.21 (Muthén & Muthén, 1998-2009). The fit of these models are summarized in Table 2.

Table 2. Model Fit for Linear and Nonlinear Models

	χ^2	*df*	*p* value	CFI	RMSEA	SRMR/WRMR
Linear model [a]						
1-factor model	201.18	14	<.001	.758	.214	0.107
Bifactor model with 1 group factor	48.07	11	<.001	.952	.107	0.040
Bifactor model with 2 group factors	14.83	8	.063	.991	.054	0.019
Bifactor model with 1 group factor and correlated errors	14.83	8	.063	.991	.054	0.019
Nonlinear model[b]						
1-factor model	126.57	7	<.001	.919	.241	1.917
Bifactor model with 1 group factor	25.84	8	.001	.988	.087	0.635

Note: CFI = Comparative Fit Index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; WRMR = weighted root mean square residual.
a. The four models listed under linear model are labeled in the text as Models 1, 2, 3, and 4, respectively.
b. The two models listed under nonlinear model are labeled in the text as Models 1 and 2, respectively.

Step 1: Assessment of CFA models. We began by fitting a one-factor model. It failed to fit adequately, $\chi^2(7) = 126.57$, $p < .01$, CFI = .92, RMSEA = .24, and weighted root mean square residual [WRMR] = 1.92. Next, Model 2, a bifactor model with the PAM group factor, was fit to the data. It yielded reasonably good fit, $\chi^2(8) = 25.84$, $p < .01$, CFI = .99, RMSEA = .09, and WRMR = .64. The factor loadings for the general factor ranged from .46 to .87, whereas the loadings for the group factor were .79, .55, and .71. The general factor accounted for approximately 89.7% of the common variance, and the scale appears to be essentially unidimensional. This value was much higher than those based on the linear models. Additional modifications showed no improvement in fit. It is possible that the negative group factor found with linear SEM methods was spurious and due to differences in thresholds between the negatively phrased items and the positively phrased items.

Step 2: Estimation of the reliability coefficient. The nonlinear SEM estimate of reliability based on Model 2 was .868. This coefficient is slightly higher than the coefficient of .860 for Model 2 and slightly lower than the coefficient of .876 for Model 3, assuming a linear relationship between items and factors. We are not likely to make any differential interpretation among the values for these three coefficients. A more important interpretational distinction between the linear and nonlinear analyses was the inclusion or exclusion of the negative group factor.

Summary From Empirical Analyses

Coefficient alpha of .815 appears to be an underestimate of the reliability for this scale. Provocatively, the results of the cross-validated maximal split-half coefficient, the linear coefficient based on Model 3, and the nonlinear coefficient based on Model 2 were very similar: .865, .876, and .868, respectively. We also gained knowledge of our scale by conducting the CFAs in conjunction with the computation of the SEM reliability coefficients. It is important to note that none of these coefficients assess transient errors.

Conclusions

On the basis of the arguments put forward in our article, we can now rewrite the previously presented reasons for using coefficient alpha as reasons for considering alternative estimates of reliability.

1. Coefficient alpha is easy to misinterpret. It should not be construed as a coefficient of internal consistency but rather as an estimate of reliability based on the internal consistency among items. As an estimate of reliability, it can be an underestimate or overestimate depending on the degree that the tau equivalence and uncorrelated errors assumptions are violated. These assumptions can be assessed using SEM. If the sample data are relatively consistent with the assumptions underlying coefficient alpha, SEM estimates of reliability should yield similar results to coefficient alpha. To the extent that the sample data indicate these assumptions are violated, it would be judicious to choose an SEM estimate of reliability that is consistent with the item data.
2. Although it is true that coefficient alpha can be computed without making subjective decisions, it is inappropriate to do so. Researchers should conduct SEM analyses (or alternative methods) to understand the structure of a scale and to assess the assumptions underlying coefficient alpha. These analyses involve subjective judgments. In addition, subjective judgment is required to assess the degree that transient errors are relevant in the application of a scale and then to choose a method to assess reliability that is consistent with that assessment (e.g., equivalent-forms reliability with a 1-week interval between administrations).
3. Coefficient alpha can be misleading if it is used to make revisions to scales. Items that are eliminated based on their effect on coefficient alpha can contribute substantially to the overall psychometric quality of a scale. More complex approaches are required to revise scales, and these more complex approaches require subjective judgments.
4. We have no doubt that researchers have developed a normative framework for interpreting values for coefficient alpha. These norms can be based on meta-analyses of previously reported alphas for a particular scale or type of scale (Hogan, Benjamin, & Brezinski, 2000; Peterson, 1994) or researchers' memory of reported alphas in their research domain. We suspect that this normative knowledge could be helpful in understanding reliability estimates based on alternative methods. In some sense, it depends on the degree that these alternative methods yield values that have similar distributions to those for coefficient alphas. Once researchers choose estimates of reliability based on careful analyses of data rather than routinely calculating coefficient alpha, we will know how right or wrong psychological researchers have been in reporting coefficient alpha.

We argue that a modeling approach to assessing reliability will be more informative than simply computing coefficient alpha. Although it requires a series of judgments by researchers, the result is a more interpretable coefficient and its underlying assumptions better understood by researchers.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Notes

1. The present discussion assumes Type 1 sampling. With this sampling approach, reliability is assessed on a scale with a fixed set of items, but individuals are randomly sampled from a population. In contrast, with Type 12 sampling, reliability is evaluated on a scale with randomly sampled items as well as randomly sampled individuals (Lord, 1955). Although it would appear that most psychometricians are assuming Type I sampling when discussing coefficient alpha, others have argued for the appropriateness of Type 12 sampling. For provocative discussions of this topic, see Barchard and Hakstian (1997) and Cronbach (2004).
2. Strictly speaking, the assumption is essential tau equivalence (Novick & Lewis, 1967). The "essential" modifier implies that any two items can differ by a constant. This distinction is not crucial for issues discussed in this article.
3. Due to space limitations, we restricted our discussion to reliability coefficients that are embedded within a classical true score framework. Generalizability theory and item response theory offer alternative frameworks for assessing reliability but are not considered in our article.
4. A number of bifactor models were fit to the data. We included only a subset of these to illustrate the stepwise process without making the presentation overly complex and tedious.

References

Barchard, K. A., & Hakstian, A. R. (1997). The effects of sampling model on inference with coefficient alpha. *Educational and Psychological Measurement, 57,* 893-905.

Becker, G. (2000). How important is transient error in estimating reliability? Going beyond simulation studies. *Psychological Methods, 5,* 370-379.

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika, 74(1)*, 137-143.

Callendar, J. C., & Osburn, H. G. (1977a). A method for maximizing split-half reliability coefficients. *Educational and Psychological Measurement, 37,* 819-825.

Callendar, J. C., & Osburn, H. G. (1977b). A computer program for maximizing and cross-validating split-half reliability coefficients. *Educational and Psychological Measurement, 37,* 787-789.

Callendar, J. C., & Osburn, H. G. (1979). An empirical comparison of coefficient alpha, Guttman's lambda2, and msplit maximized split-half reliability estimates. *Journal of Educational Measurement, 16,* 89-99.

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.

Cronbach, L. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64,* 391-418.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions*. Urbana: University of Illinois Press.

Embretson, S. E., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

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

Finney, S., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 269-314). Greenwich, CT: Information Age.

Fleishman, J., & Benson, J. (1987). Using LISREL to evaluate measurement models and scale reliability. *Educational and Psychological Measurement, 47,* 925-939.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9,* 466-491.

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference, 11.0 update* (4th ed.). Boston, MA: Allyn & Bacon.

Green, S. B. (1983). Identifiability of spurious factors using linear factor analysis with binary items. *Applied Psychological Measurement,* 7(2), 139-147.

Green, S. B. (2003). A coefficient alpha for test-retest data. *Psychological Methods, 8,* 88-101.

Green, S. B., Akey, T. M., Fleming, K. K., Hershberger, S. L., & Marquis, J. G. (1997). Effect of the number of scale points on chi-square fit indices in confirmatory factor analysis. *Structural Equation Modeling, 4,* 108-120.

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., & Thompson, M. S. (2003). Structural equation modeling in clinical research. In M. C. Roberts & S. S. Illardi (Eds.), *Methods of research in clinical psychology: A handbook.* (pp. 138-175). London, UK: Blackwell.

Green, S. B., & Yang, Y. (2005). *K*-split coefficient alpha. Paper presented at the Annual Meeting of American Education Research Association, Montreal, Québec, Canada.

Green, S. B., & Yang, Y. (2009a). Commentary on coefficient alpha: A cautionary tale. *Psychometrika, 74(1)*, 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.

Guttman, L. A. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*, 255-282.

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement, 60*, 523-531.

Hunter, J. E., & Schmidt, F. L. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*, 199-223.

Kaplan, R. M., & Saccuzzo, D. P. (2009). *Psychological testing: Principles, applications, and issues* (7th ed.). Belmont, CA: Thompson Wadsworth.

Komaroff, E. (1997). Effect of simultaneous violations of essential τ-equivalence and uncorrelated error on coefficient α. *Applied Psychological Measurement, 21,* 337-348.

Lei, P.-W. (2009). Evaluating estimation methods for ordinal data in structural equation modeling. *Quality & Quantity, 43,* 495-507.

Lord, F. M. (1955). Sampling fluctuations resulting from the sampling of test items. *Psychometrika, 20,* 1-22.

Maxwell, A. E. (1968). The effect of correlated errors on estimates of reliability coefficients. *Educational and Psychological Measurement,* 28, 803-811.

McDonald, R. P. (1978). Generalizability in factorable domains: "Domain validity and generalizability": 1. *Educational and Psychological Measurement, 38,* 75-79.

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

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

Muthén, B., & Kaplan, D. (1992). A comparison of some methodologies for the factor-analysis of non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology, 45,* 19-30.

Muthén, L. K., & Muthén, B. O. (1998-2009). *Mplus user's guide* (5th ed.). Los Angeles, CA: Authors.

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika, 32,* 1-13.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods, 5,* 343-355.

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

Raju, N. S. (1970). New formula for estimating total test reliability from parts of unequal lengths. *Proceedings of the Annual Convention of the American Psychological Association, 5,* 143-144.

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

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

Raykov, T. (1998). Coefficient alpha and composite reliability with interrelated nonhomogeneous items. *Applied Psychological Measurement, 22,* 375-385.

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,* 195-212.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16,* 19-31.

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment, 12,* 287-297.

Revelle, W., & Zinbarg, R. E. (2009). Coefficient alpha, beta, omega, and the GLB: Comments on Sijtsma. *Psychometrika, 74,* 145-154.

Rozeboom, W. W. (1966). *Foundations of the theory of prediction.* Homewood, IL: Dorsey.

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

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74(1),* 107-120.

Steinberg, L. (2001). The consequences of pairing questions: Context effects in personality measurement. *Journal of Personality and Social Psychology, 81,* 332-342.

Steinberg, L., & Thissen, D. (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods, 1,* 81-97.

Thompson, B. L., Green, S. B., & Yang, Y. (2010). Assessment of the maximal split-half coefficient to estimate reliability. *Educational and Psychological Measurement, 70,* 232-251.

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: The case for testlets. *Journal of Educational Measurement, 24,* 189-205.

Yang, Y., & Green, S. B. (2010a). A note on structural equation modeling estimates of reliability. *Structural Equation Modeling, 17,* 66-81.

Yang, Y., & Green, S. B. (2010b, May). *Evaluation of nonlinear SEM reliability coefficients.* Paper presented at Annual Meeting of the American Educational Research Association, Denver, CO.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-214.

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

Bios

Yanyun Yang is an assistant professor in the Department of Educational Psychology and Learning Systems at the Florida State University. She received Ph.D. in Educational Psychology with a concentration in Measurement, Statistics, and Methodological Studies from Arizona State University in 2007. Her research interests include structural equation modeling, psychometrics, reliability theory, and applied statistics in psychology and education.

Sam Green received his Ph.D. in Psychology, with an emphasis in Measurement and Human Differences, from the University of Georgia in 1974. He has been a faculty member at Auburn University, University of Kansas, and currently Arizona State University. His research interests include structural equation modeling, factor analysis, and reliability theory.