Psychological Methods
2016, Vol. 21, No. 1, 69–92

© 2016 American Psychological Association
1082-989X/16/$12.00 http://dx.doi.org/10.1037/a0040086

Confidence Intervals for Population Reliability Coefficients: Evaluation of Methods, Recommendations, and Software for Composite Measures

Ken Kelley
University of Notre Dame

Sunthud Pornprasertmanit
Texas Tech University

A composite score is the sum of a set of components. For example, a total test score can be defined as the sum of the individual items. The reliability of composite scores is of interest in a wide variety of contexts due to their widespread use and applicability to many disciplines. The psychometric literature has devoted considerable time to discussing how to best estimate the population reliability value. However, all point estimates of a reliability coefficient fail to convey the uncertainty associated with the estimate as it estimates the population value. Correspondingly, a confidence interval is recommended to convey the uncertainty with which the population value of the reliability coefficient has been estimated. However, many confidence interval methods for bracketing the population reliability coefficient exist and it is not clear which method is most appropriate in general or in a variety of specific circumstances. We evaluate these confidence interval methods for 4 reliability coefficients (coefficient alpha, coefficient omega, hierarchical omega, and categorical omega) under a variety of conditions with 3 large-scale Monte Carlo simulation studies. Our findings lead us to generally recommend bootstrap confidence intervals for hierarchical omega for continuous items and categorical omega for categorical items. All of the methods we discuss are implemented in the freely available R language and environment via the MBESS package.

Keywords: reliability, confidence intervals, composite score, homogeneous test, measurement

Supplemental materials: http://dx.doi.org/10.1037/a0040086.supp

Composite scores are widely used in many disciplines (e.g., psychology, education, sociology, management, medicine) for a variety of purposes (e.g., attitudinal measures, quantifying personality traits, personnel selection, performance assessment). A composite score is a derived score that is the result of adding component scores (e.g., items, measures, self-reports), possibly with different weights, in order to obtain a single value that purportedly represents some construct.[1] It can be argued that composite scores are generally most useful when the individual components are each different measures of a single construct. Measurement instruments that measure a single construct are termed homogeneous. We focus exclusively on homogeneous measurement instruments throughout the article.

Understanding various properties of a measurement instrument as it applies to a particular population is important. One property that is necessary to consider when using composite scores is the estimate of the population value of reliability for the population of interest. Although we formalize the definition of composite reliability later, recall that the psychometric definition of reliability is the ratio of the true variance to the total variance. However, it is clear that a point estimate of reliability does not go far enough, as the estimate will almost certainly not equal the population value it estimates. The value of the reliability for the population of interest, not the particular sample from the population, is what is ultimately of interest.

In practice, the population reliability coefficient is almost always unobtainable. Correspondingly, we believe that a confidence interval for the population reliability coefficient should always accompany an estimate. This statement is consistent with the American Psychological Association (American Psychological Association, 2010), American Educational Research Association (Task Force on Reporting of Research Methods in AERA Publications, 2006), and Association for Psychological Science (Association for Psychological Science, 2014) publishing guidelines, among others. A confidence interval quantifies the uncertainty of the estimated parameter with an interval, where, assuming the appropriate assumptions are satisfied, the confidence interval provides lower and upper limits that form an interval in which the population value will be contained within with the specified level of confidence. The confidence interval limits can be informally conceptualized as the range of plausible parameter values at the stated

Ken Kelley, Department of Management–Analytics, University of Notre Dame; Sunthud Pornprasertmanit, Institute for Measurement, Methodology, Analysis, and Policy, Texas Tech University.

Sunthud Pornprasertmanit is now at the Chulalongkorn University, Bangkok, Thailand. Part of this research was presented at the 2011 Joint Statistics Meeting in Miami, Florida. This research benefited from the Center for Research Computing at the University of Notre Dame and from a grant to the first author from the Notre Dame Deloitte Center for Ethical Leadership.

Correspondence concerning this article should be addressed to Ken Kelley, Department of Management–Analytics, Mendoza College of Business, University of Notre Dame, Notre Dame, IN 46556. E-mail: KKelley@ND.edu

[1] Often the component scores are added with unit weights, which is often termed unweighted. However, that need not be the case. For example, the weights of components can be any real value (e.g., -1 for reverse coded items, .5 for two items when the mean of those two items is taken, items multiplied by the estimated factor loadings to form a factor score).

confidence level.[2] Unfortunately, even if a researcher wishes to follow the established guidelines and provide a confidence interval for an estimated reliability coefficient, there are not unambiguous "best methods" for confidence interval construction for reliability coefficients.

Due to its ubiquity and importance in research and practice, the reliability of a homogeneous measurement instrument has been the subject of much attention in the psychometric literature (e.g., see Sijtsma, 2009, for a discussion with references to important historical developments and the commentaries that follow; e.g., Bentler, 2009; Green & Yang, 2009a; 2009b; Revelle & Zinbarg, 2009). However, most of the attention given to the reliability of a homogeneous measurement instrument centers on the point estimate of the population value. Feldt (1965) was the first to note the paucity of work that considered the sampling characteristics of reliability coefficients. Feldt stated that "test manuals rarely, if ever, report confidence intervals for reliability coefficients, even for a specific population" (p. 357). Unfortunately, 50 years later, confidence intervals for population reliability coefficients are often still not reported, even for composite scores with important uses.

We begin the rest of this article with a brief review of ideas relevant to classical test theory as they pertain to composite reliability coefficients. We then present four reliability coefficients before discussing confidence interval estimation for the population reliability. We then evaluate the effectiveness of the different confidence interval methods for the different reliability coefficients in a variety of contexts with three Monte Carlo simulation studies. We then make recommendations to researchers about which confidence interval methods should be used in which situations. Additionally, we provide open source and freely available software to implement each of the methods we discuss.

Classical Test Theory and Reliability Coefficients

In classical test theory (CTT) an observed item X_{ij} for the ith individual ($i = 1, \ldots, N$) on the jth component (e.g., item) ($j = 1, \ldots, J$) is decomposed into two parts as

$$X_{ij} = T_{ij} + \epsilon_{ij}, \tag{1}$$

where T_{ij} (capital tau) is the true-score for the ith individual on the jth component, and ϵ_{ij} is the error for the ith individual on the jth component (e.g., Guilford, 1954; Gulliksen, 1950; Lord & Novick, 1968; McDonald, 1999; Zimmerman, 1975). For now we will assume that the J items are continuous (but relax this assumption in a future section). The theorem in which CTT is derived states that the errors of measurement (i.e., the $\epsilon_{\cdot j}$s) are (a) mutually uncorrelated (i.e., $\rho(\epsilon_{\cdot j}, \epsilon_{\cdot j'}) = 0$ for all $j \neq j'$); (b) are uncorrelated with their corresponding true-scores (i.e., $\rho(T_{\cdot j}, \epsilon_{\cdot j}) = 0$); (c) are uncorrelated with other true scores (i.e., $\rho(T_{\cdot j'},\epsilon_{\cdot j}) = 0$ for all $j \neq j'$); and (d) have a mean of zero (i.e., $\mathrm{E}[\epsilon_{\cdot j}] = 0$), where a centered dot in place of i in the subscript denotes across individuals (Lord & Novick, 1968, Theorem 2.7.1, p. 36, see also p. 38).

A composite score is the sum of the J individual component scores as

$$Y_i = \sum_{j=1}^{J} X_{ij}, \tag{2}$$

where Y_i is the observed composite score for individual i. As is common, we do not use weights when forming the composite from the J components. More formally a composite score of the form given in Equation 2 could be called a unit-weighted composites (due to the weights of the J X values implicitly being 1). The value of Y_i is an observed score and can be conceptualized in an analogous manner as was done for X_i from Equation 1. That is, Y_i can be denoted as

$$Y_i = T_i + \epsilon_i, \tag{3}$$

where

$$T_i = \sum_{j=1}^{J} T_{ij} \tag{4}$$

and

$$\epsilon_i = \sum_{j=1}^{J} \epsilon_{ij}. \tag{5}$$

Note that there are no j subscripts for Y_i, T_i, or ϵ_i in Equation 3 because these values represent the observed composite score, the true composite score, and error for the composite score, respectively, for the ith individual.

The general representation of reliability from a psychometric perspective is the ratio of the true variance to the total variance, which for the population can be formally written as

$$\rho(Y) = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_\epsilon^2} \tag{6}$$

and rewritten as

$$\rho(Y) = \frac{\sigma_T^2}{\sigma_Y^2}, \tag{7}$$

where $\rho(Y)$ is the population reliability of the composite measure Y, σ_T^2 is the population variance of the true scores for the composite across individuals, σ_ϵ^2 is the population variance of the error of the scores for the composite across individuals, and σ_Y^2 is the population variance of the observed scores for the composite across individuals. Because T and ϵ are uncorrelated, based on the theorem defining CTT, $\sigma_Y^2 = \sigma_T^2 + \sigma_\epsilon^2$. The estimation of true and error variances (i.e., the components in Equation 6) in order to find the ratio of true-to-total variance (i.e., Equation 7) in the context of the reliability of a composite has been based on various methods in the literature (e.g., Cronbach, 1951; Guttman, 1945; Hoyt, 1941; Kuder & Richardson, 1937; McDonald, 1999; van Zyl, Neudecker, & Nel, 2000; Woodhouse & Jackson, 1977; see Raykov, 2012, for a review).

Coefficient Alpha

The most widely used reliability coefficient used to estimate $\rho(Y)$ is coefficient alpha (Cronbach, 1951). The definition of population coefficient alpha has long been given as

[2] More formally, confidence, in the context of confidence intervals, refers to the procedure that is used to form such intervals. We use C to denote the specified level of confidence (e.g., .90, .95, .99). Further, we use C100% to denote the confidence level in percentage form. For example, C100% represents a 95% confidence interval when $C = .95$. More formally, a confidence interval is an interval from a procedure in which, if it were replicated an infinite number of times, would produce intervals in which C100% of them would bracket the population value, provided that the confidence interval procedure is exact and the assumptions upon which it depends are satisfied.

$$\alpha \equiv \left(\frac{J}{J-1}\right)\left(1-\frac{\sum_{j=1}^{J}\sigma_j^2}{\sigma_Y^2}\right), \tag{8}$$

where σ_j^2 is the population variance of the jth item. The estimated value of coefficient alpha is

$$\hat{\alpha} = \left(\frac{J}{J-1}\right)\left(1-\frac{\sum_{j=1}^{J}s_j^2}{s_Y^2}\right), \tag{9}$$

When uncorrelated errors hold, as prescribed by the model defining classical test theory, the population value of coefficient alpha is equal to the population value of reliability when true-score equivalence, also termed essential tau-equivalence, is satisfied. True-score equivalence is a property in which the covariances of the J distinct items are the same but with different variances and potentially with different means (McDonald, 1999, pp. 85–86). In true-score equivalence situations each of the items are said to be equally sensitive at measuring the construct, an idea we return to in more detail momentarily. When true-score equivalence does not hold and errors are uncorrelated, the sample value of coefficient alpha is not estimating population reliability coefficient, but rather it estimates population coefficient alpha, which itself will be less than the population reliability coefficient (e.g., Novick & Lewis, 1967). In practice, we believe that relatively few measurement instruments are based on items that measure the underlying factor equally well. Correspondingly, coefficient alpha, although historically the most commonly reported estimate of the reliability of a set of scores, has major deficiencies (e.g., McDonald, 1999; Raykov & Marcoulides, 2011; Revelle & Zinbarg, 2009; Sijtsma, 2009).[3]

Coefficient Omega

McDonald (1999) approached the reliability of a homogeneous measurement instrument from a factor analytic perspective, in which a single-factor model is used to decompose the true and error variances (see also Green & Hershberger, 2000; Jöoreskog, 1971; Kelley & Cheng, 2012; Miller, 1995; Raykov & Shrout, 2002). From a factor analytic conceptualization of items, Figure 1 shows the structure of seven hypothetical items from a homogeneous measurement instrument, where ψ_j^2 is the variance of the errors of the jth item. Figure 1 is a visual representation of decomposing the true score (i.e., T_{ij}) from Equation 1 as

$$T_{ij} = \mu_j + \lambda_j \eta_i, \tag{10}$$

where μ_j is the population mean of the jth item, λ_j is the factor loading for the jth item and η_i is the factor score for the ith individual. The mean of η, the factor, is fixed to 0 and the variance of η is fixed to 1 for model identification purposes. A model of the form of Equation 10 represents a congeneric scale, which is a special type of homogeneous measurement instrument in which the items each potentially have different sensitivities and error variances that can be heterogeneous (i.e., not restricted to the same value). The factor loadings are conceptualized as the "sensitivity" of an item, which quantifies how well the items measure the factor. For a true-score equivalent situation, the factor loadings are each the same value (i.e., there is no subscript for the factor loadings).

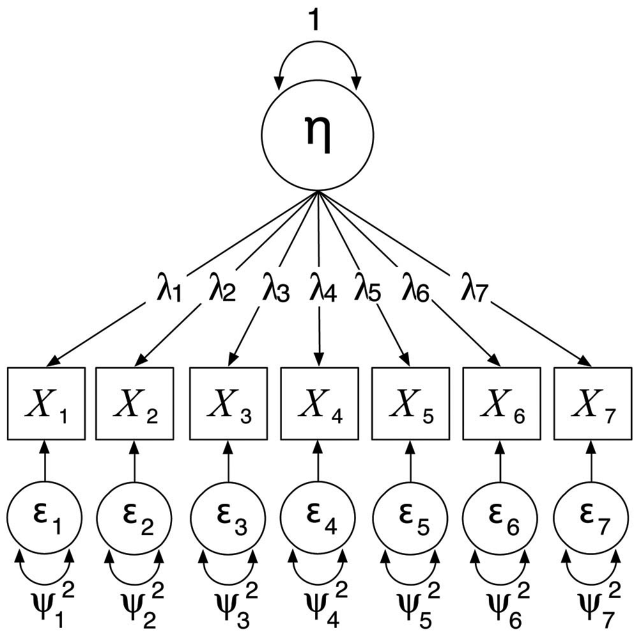

Figure 1. Path diagram of a hypothetical congeneric factor model with seven items. Note that each of the factor loadings, like the error variances, are potentially different, as they include a subscript for each item.

In this factor analytic perspective the ratio of true-to-total variance (i.e., reliability) can be shown (see Appendix) to equal

$$\omega = \frac{\left(\sum_{j=1}^{J}\lambda_j\right)^2}{\left(\sum_{j=1}^{J}\lambda_j\right)^2 + \sum_{j=1}^{J}\psi_j^2} \tag{11}$$

(see McDonald, 1999, p. 89, for more details, as well as the Appendix). For a set of items that are homogeneous with uncorrelated errors ω and the population reliability coefficient are equivalent:

$$\omega \equiv \rho(Y). \tag{12}$$

Substituting estimates (e.g., from maximum likelihood estimation) of the population values above yields an estimate of the population coefficient omega:

$$\hat{\omega} = \frac{\left(\sum_{j=1}^{J}\hat{\lambda}_j\right)^2}{\left(\sum_{j=1}^{J}\hat{\lambda}_j\right)^2 + \sum_{j=1}^{J}\hat{\psi}_j^2}, \tag{13}$$

where a circumflex above a parameter value denotes an estimate of the population value. As has been clearly articulated in the psychometric literature, the estimated coefficient omega is preferred over the estimated coefficient alpha for estimating $\rho(Y)$ because, even if errors are uncorrelated and the measurement instrument is homogeneous, unless true-score equivalence holds (i.e., the factor loadings from a single factor model are the same), the estimated

[3] When uncorrelated errors does not hold coefficient alpha may overestimate the value of the population reliability (Green & Hershberger, 2000; Komaroff, 1997; Zimmerman, Zumbo, & Lalonde, 1993).

coefficient alpha will tend to underestimate the population reliability (i.e., $\alpha \leq \rho(Y)$; thus $\mathrm{E}[\hat{\alpha}] \leq \rho(Y)$) (McDonald, 1999, p. 92).

Importantly, we want to emphasize that each λ and each ψ^2 in Figure 1 have j subscripts, which acknowledges that in this framework (for coefficient omega) the error variances (as is true for coefficient alpha) can vary by item but so too can the factor loadings. Thus, when estimating reliability from the coefficient omega perspective there is no assumption that each item is equally sensitive at measuring the factor. A scale of the general form given in Figure 1 and as discussed here, where the λs can be unique (i.e., there are j such values) is termed congeneric. This differs from our discussion of true-score equivalence, in that true-score equivalence requires that the items measure the construct with equal sensitivity. In our experience, the assumption of true-score equivalence (i.e., equally sensitive items) is untenable in many situations in which a composite score is of interest.

Hierarchical Omega

All items in a homogeneous scale are manifestations presumed to be caused by the same factor. One implication of a homogeneous scale is that, after partialing out the influence of the common factor, all items are independent (recall that classical test theory has this as part of the defining theorem). Implicit in a homogeneous scale is that there are no subdomains or facets among the items. In practice, however, it is difficult to have perfectly independent items after partialing out the variance due to the factor. For example, there may be some pairs of items that are correlated for reasons beyond the factor itself, such as similarly worded items, the content of the items overlap an additional construct, extraneous knowledge is required for a subset of items, et cetera. We will refer to these influences among some items above the factor itself as "minor factors." In such situations an observed item is decomposed into more than only the factor and the error that we have discussed thus far. Instead, the jth component for the ith individual is decomposed as

$$X_{ij} = \mu_j + \tau_{Cij} + \tau_{1ij} + \cdots + \tau_{Mij} + \epsilon_{ij}, \tag{14}$$

where τ_{Cij} is the part of true score from the common factor for the ith individual on the jth component and τ_{mij} is the part of true score from the mth minor factor for the ith individual on the jth component.

Although Equation 14 represents nonhomogeneous items, a researcher may assume that minor factors are ignorable and treat the items as if they are explained by a homogeneous (i.e., single factor) model. Minor factors can be reparameterized as correlations among the residuals of the relevant items compromising the minor factor to avoid model misspecification. If a scale with minor factors exists and it is fitted with a single factor model (i.e., under the assumption of a homogeneous set of items), the model is misspecified (because there are correlations among some items due to the minor factor(s) being ignored). Failure to appropriately address minor factors among a set of items yields a misspecified model that, as a consequence, will lead to the sum of the true variance and error variance not being equal to the total variance of the composite. Correspondingly, when minor factors are present, neither α or ω equal $\rho(Y)$.

If the model-implied covariance matrix from the misspecified model is used to estimate the total variance (i.e., when minor factors exist but are ignored), the total variance as calculated from summing the true plus error variance will differ from the actual total variance of the composite (i.e., the variance of Y). If the population reliability is estimated from the total variance implied from a misspecified model in which minor factors exist, the estimated value will have an expectation that underestimates or overestimates the population value, depending on the structure of the residual covariances. Because residual covariances cannot be estimated by the factor analysis model, as the model would not be identified, the appropriate total variance cannot be estimated by model parameters from the factor analysis model. The variance of the composite can be obtained from the observed covariance matrix of the items or equivalently by the variance of Y. Using the ideas of coefficient omega but with variance of Y in the denominator of Equation 13, the population reliability coefficient calculated in such a manner we call population hierarchical omega (ω_H):

$$\omega_H \equiv \frac{\left(\sum_{j=1}^{J} \lambda_j\right)^2}{\sum_{j=1}^{J} \sum_{j'=1}^{J} \sigma_{jj'}} = \frac{\left(\sum_{j=1}^{J} \lambda_j\right)^2}{\sigma_Y^2}, \tag{15}$$

which has a sample analog calculated by replacing the population parameters with their corresponding sample values.

Researchers never know with certainty all possible minor factors that may exist in a population and allowing the set of residual to correlate cannot be done (as doing so would yield a model that is not identified). However, knowing the items that are measures of a minor factor or allowing all residuals to correlate is not needed in our approach because we use the total variance for a set of items calculated directly from the composite score (rather than obtaining a total variance from the result of a potentially misspecified model). Hierarchical omega acknowledges that such minor factors might exist and calculates the total variance directly from the variance of the composite. Hierarchical omega is more general than coefficient omega. However, the two are equal in models in which there are no minor factors (or correlations among errors).

The reason that we have named this type of reliability "hierarchical omega" is because the formula is similar to the reliability for the hierarchical factor model proposed by McDonald (1999) and Zinbarg, Yovel, Revelle, and McDonald (2006). McDonald (1999) proposed that for the hierarchical factor model the reliability should represent the squared correlation of the total test score with the hierarchical general factor. The reason is that reliability should also reflect criterion validity, the degree to which the composite score reflects the true score that it is intended to measure. Therefore, specific factors should not be used to compute the reliability because they are not intended to measure the construct of interest. The reliability in Equation 15 (i.e., hierarchical omega) serves the same purpose for reliability coefficients of a composite as the reliability coefficient from a hierarchical factor model. The difference, however, is that here the minor factors described in this section are not included in the model because they are unknown (or thought to be ignorable), whereas specific factors in McDonald (1999) are explicitly included in a model.

What we have called hierarchical omega and the way in which it accounts for the potential of minor factors, by using the observed variance of Y, serves as a link between coefficient omega and categorical omega and has not been discussed in the literature before. Categorical omega, which is discussed in the next section, uses the information from bivariate item polychoric correlations to calculate the total variance of Y. However, coefficient omega calculates total variance from the sum of true-score variance and error variance derived from model parameters. There is no direct counterpart of categorical omega that uses the observed covariance matrix to calculate the total variance of Y so our hierarchical omega serves this purpose. We also highlight that both hierarchical omega and categorical omega have advantages over coefficient omega because they do not assume that the confirmatory factor analysis (CFA) model fits perfectly.

Categorical Omega

Until now we have considered all items to be continuous measures. However, this is often an unrealistic assumption because many of the mostly widely used scales use relatively few ordered categories (e.g., *strongly disagree*, *disagree*, *neutral*, *agree*, *strongly agree*). When items are ordered-categorical, X_{ij} cannot be represented as the sum of a true score and error as shown in Equation 1 for continuous items. Rather, the sum of the true score and error represent the continuous response variable presumably underlying X_{ij}, denoted as X^*_{ij}. The relationship between X^*_{ij} and X_{ij} can be modeled with a probit link function (Muthén, 1984). In particular,

$$X_{ij} = c \quad \text{if} \quad t_{j,c} < X^*_{ij} \leq t_{j,c+1}, \tag{16}$$

where $c = 0, 1, \ldots, C - 1$ and $t_{j,c}$ represents thresholds of indicator j separating categories $c - 1$ and c, $t_{j,0} = -\infty$ and $t_{j,C} = \infty$. Setting $\mathrm{Var}(X^*_{ij}) = 1$, which is known as the delta parameterization (Millsap & Yun-Tein, 2004), the model-implied population covariances for the categorical scale score (i.e., the composite) can be calculated as follows:

$$\sigma_{jj'}(\tilde{\rho}_{jj'}) = \sum_{c=1}^{C-1}\sum_{c'=1}^{C-1} \Phi_2(t_{j,c}, t_{j',c'}, \tilde{\rho}_{jj'}) - \left(\sum_{c=1}^{C-1} \Phi_1(t_{j,c})\right)\left(\sum_{c'=1}^{C-1} \Phi_1(t_{j',c'})\right) \tag{17}$$

where $\Phi_1(t_{j,c})$ is the cumulative probability of $t_{j,c}$ given a univariate standard normal distribution and $\Phi_2(t_{j,c}, t_{j',c'}, \tilde{\rho}_{ij})$ is the joint cumulative probability of t_{jc} and $t_{j'c'}$ given a bivariate standard normal distribution with a population correlation of $\tilde{\rho}_{jj'}$.

The model-implied covariance for the categorical scale score can be used to calculate categorical reliability (Green & Yang, 2009a). The population value of categorical omega is given as

$$\omega_C = \frac{\sum_{j=1}^{J}\sum_{j'=1}^{J} \sigma_{jj'}(\lambda_j\lambda_{j'})}{\sum_{j=1}^{J}\sum_{j'=1}^{J} \sigma_{jj'}(\rho_{X^*_jX^*_{j'}})}, \tag{18}$$

where $\sum_{j=1}^{J}\sum_{j'=1}^{J} \sigma_{jj'}(\lambda_j\lambda_{j'})$ (i.e., the numerator of Equation 18) represents the variance explained by true scores and $\rho_{X^*_jX^*_{j'}}$ is the polychoric correlation between items j and j'. The estimated value of categorical omega can be obtained, we use robust weighted least square (Flora & Curran, 2004), and substituted for the population values in Equation 18.

It is useful to note that Equation 18 is equivalent to hierarchical omega because a polychoric correlation is equivalent to the observed item correlation. Polychoric correlation is the correlation between two latent variables presumed to be underlying the ordered categorical items (and not based on the model-implied item covariances as done for coefficient omega). Green and Yang (2009a) found in the situations they investigated that if a scale has both negatively and positively skewed distributed items, categorical omega has higher values than hierarchical omega when no minor factors are present. That is, controlling for the polychoric correlation, the Pearson's correlation coefficient between categorical items is lower when the skewness of both items is in different directions (see Crocker & Algina, 1986 for numerical illustration). Hence, coefficient omega underestimated categorical omega when response distributions were skewed in different directions.[4]

Methods of Confidence Interval Formation for Population Reliability Coefficients

We have now discussed four different point estimates of the population composite reliability. However, a point estimate alone does not acknowledge the uncertainty with which the population value has been estimated. The population value, not the sample value, is what is ultimately of interest. Many different methods to calculate a confidence interval for the population reliability coefficient have been proposed. In this section we briefly discuss each of five classes of confidence interval formation methods that we evaluate in the next section. We aim for our brief summaries of the confidence interval methods to provide a general overview of each of the confidence interval procedures we evaluate.

Feldt's Approach

Parallel items is a restrictive structure in which the items have equal covariances to each other and equal variances (p. 86 McDonald, 1999). From the factor analytic perspective, parallel items imply that the factor loadings are all the equal to one another and that the error variances are all equal to one another. Thus, from Figure 1 there are no j subscripts. If the item distribution is multivariate normal and the items are parallel, Feldt (1965) and Feldt, Woodruff, and Salih (1987) showed that the approximate $1 - \hat{\alpha}$ confidence interval for coefficient alpha can be derived using an F-distribution as

$$\begin{aligned}\Pr[1 - (1 - \hat{\alpha})F_{(1-C)/2,N-1,(N-1)(J-1)} &\leq \rho(Y) \leq 1 \\ - (1 - \hat{\alpha})F_{1-(1-C)/2,N-1,(N-1)(J-1)}] &\approx C,\end{aligned} \tag{19}$$

where Pr() represents the probability of the bracketed terms, $\hat{\alpha}$ is the estimated value of coefficient alpha, N is sample size, J is the number of items, F_{A,df_1,df_2} is the value of the Ath quantile of the F distribution with degrees of freedom of df_1 and df_2, and C is the confidence level.[5]

[4] Raykov, Dimitrov, and Asparouhov (2010) proposed a reliability index for binary items. Because the Raykov et al. (2010) approach is for only two categories, and in practice many scales have more than two categories, we elected not to include it in our evaluation of reliability coefficients. Instead, we used the Green and Yang (2009a) categorical omega approach because of its generalizable nature, as it can be used for any number of categories.

[5] Siotani, Hayakawa, and Fujikoshi (1985) used a similar formula to Feldt's (1965)'s method. The difference is that the degrees of freedom in the F-distribution for the Siotani et al. (1985) approach are N and $N(J - 1)$, rather than $N - 1$ and $(N - 1)(J - 1)$ from Feldt's method. Because the difference between the degree of freedom values will generally be trivial, we included only Feldt's method in our simulation study. Note that the method of Siotani et al. (1985) is also referred to as Koning and Franses (2003)'s exact method in Romano et al. (2010).

Although we believe that the parallel assumption is generally unrealistic, we include this method here to make the current simulation comparable with previous simulation studies that included it (Cui & Li, 2012; Romano, Kromrey, & Hibbard, 2010). We also applied this method to coefficient omega because coefficient omega equals coefficient alpha under the parallel assumption. Thus, more generally, $\hat{\alpha}$ in Equation 19 can be replaced with $\hat{\rho}(Y)$, where $\hat{\rho}(Y)$ represents either $\hat{\alpha}$ or $\hat{\omega}$ depending on the condition.

Delta Method

The delta method is a way to obtain an approximate distribution of a function, including estimates of its moments such as the mean and variance, using an approximation of the function that is easier to deal with (e.g., Oehlert, 1992).

There are two ways to apply the delta method to estimate the standard error of coefficient alpha. First, van Zyl, Neudecker, and Nel (2000) used the delta method to derive an approximate variance of the sample coefficient alpha that assumes multivariate normality of the items for a true-score equivalent model (i.e., more lenient than the parallel items assumption as used in the Feldt approach, in that true-score equivalence allows heterogeneous error variances).[6] The variance from this approach for the sample coefficient alpha is

$$\mathrm{Var}(\hat{\alpha}) = \frac{1}{N}\frac{J^2}{(J-1)^2}\frac{2}{(\mathbf{1}'\mathbf{S}\mathbf{1})^3}\big[(\mathbf{1}'\mathbf{S}\mathbf{1})(\mathrm{tr}(\mathbf{S}^2) + \mathrm{tr}^2(\mathbf{S})) - 2\mathrm{tr}(\mathbf{S})(\mathbf{1}'\mathbf{S}^2\mathbf{1})\big], \qquad (20)$$

where $\mathbf{S}$ is a $J \times J$ sample covariance matrix among items, tr() is the trace function of a matrix (i.e., the sum of diagonal elements of the matrix), $\mathbf{1}$ is a $J \times 1$ vector in which all elements are 1, and, to be clear, $\mathbf{S}^2$ is matrix multiplication of the matrix with itself: $\mathbf{S}^2 = \mathbf{S} \times \mathbf{S}$). This method assumes that coefficient alpha is asymptotically normally distributed (i.e., the sampling distribution is normally distributed when the sample size approaches infinity) and, consequently, the confidence interval is symmetric. The Wald (normal-theory) confidence interval with confidence level C can be computed as

$$\Pr\Big[\hat{\alpha} - z_{1-(1-C)/2} \times \sqrt{\mathrm{Var}(\hat{\alpha})} \leq \alpha \leq \hat{\alpha} + z_{1-(1-C)/2} \times \sqrt{\mathrm{Var}(\hat{\alpha})}\Big] \approx C, \qquad (21)$$

where $z_{1-(1-C)/2}$ is the $1 - (1 - C)/2$ quantile of the standard normal distribution.[7] Note that the confidence interval obtained from Equation 21 is a Wald confidence interval, in that this interval, like others to follow, is formed as an estimate plus-and-minus the product of the appropriate critical value and standard error.

Second, Maydeu-Olivares, Coffman, and Hartmann (2007) used the asymptotically distribution free method (ADF) to derive the standard error of coefficient alpha. However, unlike the normal-theory approach, the ADF approach does not require the observed random variables to be normally distributed because the ADF approach uses the variance and the shape of the distributions of items to find the standard error of the coefficient alpha (Browne, 1984). The minimizing function is based on the asymptotic covariance matrix of the unique elements of the item covariance matrix. That is, the unique elements of the item covariance matrix are stacked. Then, the asymptotic covariance matrix of the unique elements is calculated, in which the square root of each diagonal element is the corresponding standard error. This asymptotic covariance matrix is used as the weight in the minimizing function and used to calculate the standard error of coefficient alpha. The standard error of coefficient alpha will be weighted more by (a) items that have smaller standard errors (more precision), and (b) items that have low nondiagonal elements of the asymptotic covariance matrix. This implies that the standard error of coefficient alpha will be weighted more by items that provide less redundant information about the composite score.

The ADF approach does, however, assume that the sampling distribution of the estimated coefficient alpha values is asymptotically normally distributed (Wald confidence interval). Consequently, the ADF approach assumes that the distribution of the coefficient alpha is symmetric. In addition, the ADF method requires the asymptotic covariance matrix of the unique elements of the item covariance matrix. The asymptotic covariance matrix of the unique elements can be quite large. As examples, the size of the covariance matrix would be 15 × 15 for computing a coefficient alpha from a five-item scale (i.e., 5 × 6/2) and 78 × 78 for a 12-item scale (i.e., 12 × 13/2). Because the number of the parameters needed to be estimated in the covariance matrix can be large, a large sample size is often necessary to yield an accurate estimate of the standard error.

For coefficient omega, the normal-theory approach or the ADF approach can be used. Raykov (2002a) used the normal-theory approach to derive a closed-form solution for the variance of coefficient omega assuming that data are multivariate normally distributed (i.e., using maximum likelihood estimation). Alternatively, nonlinear constraints can be used to find the standard error of coefficient omega when a one-factor model is estimated in structural equation modeling packages (Cheung, 2009; Raykov, 2002a). For violations of the assumption of normally distributed items, robust maximum likelihood estimation (Bentler & Satorra, 2010; Satorra & Bentler, 2001) could be used to approximate standard errors that are more robust to normality violation.

For the ADF approach, a closed-form equation for the standard error is not available to our knowledge. However, with nonlinear constraints in SEM programs, ADF can be implemented for the estimate of coefficient omega and its standard error. The difference is the change of the estimation method from maximum likelihood estimation to ADF. Using the ADF estimation method, however, the parameter estimates of coefficient omega are estimated by a different discrepancy function from the maximum likelihood method, which involves the asymptotic covariance matrix of the unique elements of item covariances. As noted in the context of coefficient alpha, the ADF estimation method can be advantageous because it is not based on the multivariate normality assumption of items (Browne, 1984; Olsson, Foss, Troye, & Howell, 2000). However, the obtained point estimate might be biased (Olsson et

[6] van Zyl et al. (2000) also proposed the reduced form of this formula when the items are parallel. The parallel-form formula was referred to as "Koning and Franses's approximate method" in Romano et al. (2010). Note also that the method of van Zyl et al. (2000) was referred to by Romano et al. (2010) as Iacobucci and Duhachek (2003)'s method.

[7] If Type I error is denoted as α', the quantile of interest from the standard normal distribution could be denoted as $z_{1-\alpha'/2}$ rather than $z_{1-(1-C)/2}$. In an effort not to confuse the Type I error rate with coefficient alpha, we use C for the confidence interval coefficient. It will always be the case that $\alpha' + C = 1$.

al., 2000) in small sample size situations because of the inaccuracy in the asymptotic covariance matrix. Therefore, if the confidence interval procedure fails to perform well, it may be due to the method yielding biased point estimates of the population value of coefficient omega or due to the standard error being inappropriate. The limit of the confidence interval for coefficient omega from both the normal-theory and ADF approaches is that the confidence intervals are calculated in the form of Wald confidence intervals (i.e., symmetric confidence intervals).

Transformation-Based Approaches

The delta method does not acknowledge that, when sample size is small, the reliability coefficient (e.g., $\hat{\alpha}$ or $\hat{\omega}$) is rarely normally distributed. Because the scale of reliability can range from 0 to 1, the margin of error on the side closer to a limiting bound tends to be shorter than on the other side.[8] Taking a population reliability coefficient of .99 as an example, the sampling distribution of an estimated reliability coefficient in finite samples will have a nontrivial proportion of the estimated reliability coefficients being smaller than the population value but due to the boundary at 1.0 (i.e., perfect reliability), no possibility of having an estimated value larger than 1.0. Thus, comparing the proportion of values from the sampling distribution less than .99 to the proportion greater than .99 will yield an asymmetry (a much higher proportion of sample values will be less than the population value here). Thus, in such a situation, the sampling distribution is negatively skewed.

The transformation-based approach acknowledges this potential asymmetry by transforming the estimated reliability coefficient so that the new different scale has a sampling distribution closer to normality, even in a small sample. Based on the result of previous simulation studies (e.g., Padilla, Divers, & Newton, 2012; Romano et al., 2010), we included Fisher's, Bonett's, Hakstian and Whalen's, and logistic transformations in this study for both coefficient alpha and coefficient omega. Romano, Kromrey, and Hibbard (2010) found that Fisher's and Bonett's methods performed better than the alternatives they investigated with regards to the nominal and empirical confidence interval coverage. Padilla, Divers, and Newton (2012), however, found that Fisher's method was not ideal because it had high coverage rates (i.e., 98% or higher when the nominal coverage was 95%). Bonett's method, however, performed well in most situations.

The computation of the confidence interval based on the Fisher's z' transformation requires several steps (Fisher, 1950).[9] First, the estimated value of the population reliability coefficient, generically denoted $\hat{\rho}$ for whatever estimate one chooses (e.g., coefficient alpha or coefficient omega), is transformed to the z' scale by

$$\hat{z}' = \frac{1}{2}\log\left(\frac{1+\hat{\rho}}{1-\hat{\rho}}\right), \tag{22}$$

where $\hat{\rho}$ is an estimate of the population sample reliability coefficient and log() is the natural logarithm function. Second, the confidence interval for the transformed variable is then formed as

$$\Pr\left[\hat{z}' - z_{1-(1-C)/2}\sqrt{\frac{1}{N-3}} \leq z' \leq \hat{z}' + z_{1-(1-C)/2}\sqrt{\frac{1}{N-3}}\right] \approx C, \tag{23}$$

where z' is the population value of the transformed reliability coefficient. The upper and lower bounds of the confidence interval for the z' (i.e., the transformed confidence interval) are transformed into the original reliability metric so that the limits are on the appropriate scale. In particular, the upper and lower bounds of the transformed confidence interval for reliability are transformed back to the original reliability metric by

$$\hat{\rho} = \frac{\exp(2\hat{z}') - 1}{\exp(2\hat{z}') + 1}. \tag{24}$$

The method of Bonett (2002) is based on the transformation of the intraclass correlation coefficient.[10] Bonett's transformation formula is

$$\hat{z}'' = \log(1 - \hat{\rho}), \tag{25}$$

where a double prime superscript is used to distinguish it from Fisher's transformation.[11] The confidence interval for the Bonett's transformed reliability is

$$\Pr\left[\hat{z}'' - z_{1-(1-C)/2}\sqrt{\frac{2J}{(J-1)(N-2)}} \leq z'' \leq \hat{z}'' + z_{1-(1-C)/2}\sqrt{\frac{2J}{(J-1)(N-2)}}\right] \approx C. \tag{26}$$

Similar to the Fisher's approach, the upper and lower bounds of the transformed confidence interval are transformed back into the original reliability metric.[12]

$$\hat{\rho} = 1 - \exp(\hat{z}''). \tag{27}$$

Hakstian and Whalen's (1976) approach to confidence intervals for reliability coefficients uses the cube-root transformation of $1 - \hat{\rho}$. The Hakstian and Whalen's transformation formula is

$$\hat{z}''' = (1 - \hat{\rho})^{1/3}, \tag{28}$$

[8] Robust maximum likelihood estimation and ADF do not assume that the item distribution is multivariate normal. However, the resulting confidence interval assumes that the sampling distribution of estimated reliability coefficients is normally distributed. That is, it is important to realize that there are two types of normality in this context.

[9] Fisher proposed this z' transformation in order to obtain a better confidence interval for the population Pearson's product-moment correlation, yet it has been adapted to the context of reliability. Reliability can be interpreted as the correlation between a composite score and another measure of the composite score (e.g., McDonald, 1999, p. 66). Thus, using a correlation transformation for a reliability coefficient does have a solid grounding. However, unlike correlation, reliability cannot take negative values.

[10] Reliability is the proportion of observed score variance that is attributed to the variation of true score (Crocker & Algina, 1986), which is a concept similar to intraclass correlation.

[11] There are several proposed formulas similar to Bonett (2002) formula. The first variation is that the standard error of the transformed variable is $2J/[N(J - 1)]$ (e.g., Fisher, 1950; van Zyl et al., 2000). The second variation is that the transformation formula is $\log(1 - \hat{\rho}) - \log[N/(N - 1)]$ (Bonett, 2010). These are within the same family of transformations. Therefore, we consider only the Bonett (2002) formula.

[12] Bonett (2002) noted that "unless computational ease is a primary concern, the exact confidence interval (Feldt et al., 1987) would be used instead" of the proposal Bonett (2002) offered (p. 337). Interestingly, previous simulation studies revealed that Bonett's approach worked better than the Feldt's approach in a variety of situations (Cui & Li, 2012; Padilla et al., 2012; Romano et al., 2010).

where a triple prime superscript is used to distinguish it from Fisher's and Bonett's transformations. The confidence interval for the Hakstian and Whalen's transformed reliability is

$$\Pr\left[\hat{z}''' - z_{1-(1-C)/2}\sqrt{\frac{\frac{2J}{9(n-1)(J-1)}(1-\hat{\rho})^{2/3}}{\left(1-\frac{2}{9(n-1)}\right)^2}} \le z''' \le \hat{z}''' + z_{1-(1-C)/2}\sqrt{\frac{\frac{2J}{9(n-1)(J-1)}(1-\hat{\rho})^{2/3}}{\left(1-\frac{2}{9(n-1)}\right)^2}}\right] \approx C. \quad (29)$$

Similar to Fisher's and Bonett's approaches, the upper and lower bounds of the transformed confidence interval for reliability are transformed into the original reliability metric with the following transformation

$$\hat{\rho} = 1 - \left(z'''\frac{1-\frac{2}{9(n-1)}}{1-\frac{2}{9(n-1)(J-1)}}\right)^3. \quad (30)$$

Finally, because reliability ranges from 0 to 1, the logistic link can expand the range of the limits from 0 and 1 to $-\infty$ and ∞. Unlike the previous transformations, the logistic transformation takes both point estimate and standard error of reliability estimate to find the confidence interval for population reliability. Thus, the point estimate and standard errors from the delta method can be used to build a nonsymmetic confidence interval. First, the point estimate of reliability is transformed based on logistic transformation (Browne, 1982; Raykov & Marcoulides, 2013):

$$\hat{\kappa} = \log(\hat{\rho}/(1-\hat{\rho})), \quad (31)$$

where $\hat{\kappa}$ is the logistic-transformed reliability. Then, the confidence interval for the logistic-transformed reliability is

$$\Pr\left[\hat{\kappa} - z_{1-(1-C)/2}\left(\frac{SE(\hat{\rho})}{\hat{\rho}(1-\hat{\rho})}\right) \le \kappa \le \hat{\kappa} + z_{1-(1-C)/2}\left(\frac{SE(\hat{\rho})}{\hat{\rho}(1-\hat{\rho})}\right)\right] \approx C. \quad (32)$$

Then, the confidence interval for ρ can be obtained by the inverse function of logistic transformation:

$$\hat{\rho} = 1/(1+\exp(-\hat{\kappa})). \quad (33)$$

This method does not generally produce a symmetric confidence interval, which is fine for a bounded quantity such as the reliability coefficient.

Likelihood-Based Approach

When the delta-method confidence intervals are formed as an estimate plus-and-minus a single value for the margin of error, as those above, there is an assumption of symmetry of the sampling distribution of the statistic. A way to avoid the assumption of symmetry is using what can be termed the *shifting method*. The confidence interval by the shifting method can be formed by several steps: (a) the point estimate of reliability ($\hat{\rho}$) is calculated; (b) a hypothesized population reliability value is arbitrarily made at the point where the estimated reliability is posited; (c) the hypothesized reliability value is shifted to smaller and smaller values until the appropriate test statistic "turns" from being not being statistically significant to being statistically significant, with this *turning point* used as the lower bound of confidence interval; and (d) this process is repeated for the upper confidence interval limit by shifting the hypothesized reliability value to the larger value side. The distance between the lower bound and the parameter estimate may not be equal to the distance between the upper bound and the parameter estimate (i.e., the confidence interval need not be symmetric).

If the test statistic is obtained by a Wald test, the shifting method is equivalent to the delta method and the confidence interval is symmetric. However, rather than using the Wald test, the likelihood-ratio (LR) statistic for nested model comparisons is used. Let $L(\hat{\rho}(Y))$ be the likelihood of the model when reliability is estimated and $L_0(\rho(Y))$ be the likelihood of the model when reliability is fixed. Likelihoods from both models will be compared by using likelihood-ratio statistics, such as

$$\text{LR} = 2(\log[L_0(\rho(Y))] - \log[L(\hat{\rho}(Y))]). \quad (34)$$

This likelihood-ratio statistic is (asymptotically) chi-square distributed with one degree of freedom. This test statistic is used to find the upper and lower bounds of a confidence interval by finding the turning points in which the test statistic turns from not being statistically significant to being statistically significant. Some structural equation modeling packages can be used to find likelihood-based confidence intervals (we use OpenMx; Boker et al., 2011). To find the confidence interval for population reliability, nonlinear constraints are added in the model (see more details in Cheung, 2009).

Bootstrap Approaches

Bootstrap approaches posit that it is better to base confidence intervals on empirical distributions rather than theoretical distributions (Efron & Tibshirani, 1993). If the sample distribution deviates drastically from a normal distribution, provided sample size is not small, there is thus some empirical evidence that the distribution in the population differs from normality.[13] Further, it might be very clear that normality is suspect at best from a theoretical perspective. The bootstrap approach uses an empirical distribution of the estimate(s) in order to better approximate the sampling distribution of the estimate that exists in reality, as evidenced from the sample data, as opposed to a sampling distribution based on a theoretical mathematical model, such as the normal distribution. The sampling distribution of the estimate in a bootstrap context is based on the idea of sampling a large number of times (e.g., 10,000) N observations with replacement from the original scores in order to form an empirical distribution of the desired statistic(s). Then, a confidence interval can be obtained from the empirical sampling distribution from the bootstrap procedure. Bootstrap approaches seem useful for confidence intervals for reliability coefficients because, in practice, item distributions are often not likely to be normally distributed, which is an assumption normal-theory based methods (e.g., maximum likelihood) make. Some have suggested using bootstrap approaches to find confidence intervals for the reliability coefficient (e.g., Kelley &

[13] Although the assumption of normality can be evaluated formally with a significance test, we do not suggest such a two-stage procedure. See (Rasch, Kubinger, & Moder, 2011, and the references contained therein) for a discussion of pretesting assumptions for tests of statistical inference.

Cheng, 2012; Raykov, 2002b; Yuan, Guarnaccia, & Hayslip, 2003) without formally evaluating the bootstrap methods' effectiveness against alternatives. Some researchers have conducted simulation studies that showed the bootstrap approach outperformed certain alternatives (Cui & Li, 2012; Padilla & Divers, 2013a; Padilla et al., 2012; Romano et al., 2010). However, these studies did not investigate the performance of all four of the reliability coefficients we have discussed in as diverse situations. Additionally, we evaluate three bootstrap approaches for all four reliability coefficients under all of the conditions.

The first bootstrap approach is the bootstrap standard error method. The variance of the estimated reliability across the bootstrap samples is calculated. The variance of the estimate is used in a typical Wald-type confidence interval (i.e., Equation 21), where the square root of the variance of the bootstrap estimates is used for the standard error. The assumption of this approach is that the confidence interval is symmetric. However, an alternative is to use the logistic transformation applied in the bootstrap standard error method. The logistic transformation from Equation 31–33 can be applied within the bootstrap standard error method. The transformation works by transforming the reliability estimate to the logistic scale and with the estimated standard error being derived from the standard deviation across bootstrap samples of the transformed estimate. Note that this transformation does not assume that the confidence interval is symmetric.

The second bootstrap approach is the percentile method. The percentile method finds the values at the specified percentile from the empirical distribution of the statistic so that upper and lower confidence interval limits can be found. In particular, the percentile method finds the values from the empirical distribution at the $(1 - C)/2$ and $1 - (1 - C)/2$ percentiles to use as the lower and upper bounds of the confidence interval.

The third bootstrap approach is the bias-corrected and accelerated approach (BCa). The BCa is used to find a confidence interval that corrects for bias in the point estimate and considers the skewness of the sampling distribution (i.e., asymmetric sampling distributions will get an adjustment in the confidence interval). The difference between percentile and BCa approaches is that BCa uses different percentiles to provide the lower and upper bounds of a reliability coefficient. The lower and upper percentiles, $\%L_{\text{BC }a}$ and $\%U_{\text{BC}a}$, respectively, are used for the BCa confidence interval as

$$\%L_{\text{BC}a} = \Phi_1\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{(1-C)/2}}{1 - \hat{a}(\hat{z}_0 + z_{(1-C)/2})}\right) \tag{35}$$

and

$$\%U_{\text{BC}a} = \Phi_1\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{1-(1-C)/2}}{1 - \hat{a}(\hat{z}_0 + z_{1-(1-C)/2})}\right), \tag{36}$$

where $\Phi_1()$ is the standard normal cumulative distribution, $z_{(1-C)/2}$ and $z_{1-(1-C)/2}$ are critical values of the desired confidence level coverage from the standard normal distribution. The $\hat{z}_0$ can be computed by first finding the proportion of the number of bootstrap reliability estimates that are less than the observed reliability estimate and then finding the inverse standard normal cumulative distribution (Φ_1^{-1}) of the proportion from the first step. $\hat{a}$ is the skewed sampling distribution adjustment (Efron & Tibshirani, 1993) computed by

$$\hat{a} = \frac{\sum_{i=1}^{N} (\tilde{\rho}_{(.)} - \tilde{\rho}_i)^3}{6\left[\sum_{i=1}^{N} (\tilde{\rho}_{(.)} - \tilde{\rho}_i)^2\right]^{3/2}}, \tag{37}$$

where $\tilde{\rho}_i$ is the jackknife estimate of reliability when the ith individual's set of scores is removed and $\tilde{\rho}_{(.)}$ is the mean jackknife estimate of ρ, which is $(\sum_{i=1}^{N} \tilde{\rho}_i)/N$.

Monte Carlo Simulation Studies

We use our Monte Carlo simulation studies to compare the performance of the different confidence interval methods in conjunction with the four reliability coefficients we discussed. In general, there is no known way to analytically derive the actual confidence interval coverage, characteristics of the confidence interval width (e.g., mean width), or the overall properties of the estimation methods in finite sample sizes. Correspondingly, Monte Carlo simulation studies were necessary to evaluate the various confidence interval methods for the reliability coefficients under a variety of conditions that we believe represent realistic scenarios so as to be able to generalize our results to research in psychology and related disciplines.

We conduct three studies. Study 1 replicates and extends previous simulation studies evaluating the performance of confidence interval methods for coefficient alpha and coefficient omega to estimate population reliability. We found that coefficient alpha tends to underestimate the population reliability in situations where tau-equivalence does not hold and that it does not tend to bracket population reliability at the stated level of confidence. In Study 2 we study the performances of different confidence interval methods for coefficient omega and hierarchical omega in estimating population reliability. Study 2 differs from Study 1 because the model does not have a perfect fit at the population level (i.e., there is model misspecification). In Study 2 we find that hierarchical omega outperforms coefficient omega when the model fit is not perfect. In Study 3 we compare the performances of confidence interval methods of hierarchical omega and categorical omega in estimating population reliability when items are ordered categorical. We found that categorical omega was better than hierarchical omega for a scale with ordered categorical items. Table 1 provides an overview of the factors investigated in the three simulation studies (which we detail more fully when we discuss the methods of each study).

All confidence intervals were computed using the `ci.reliability()` function (Kelley & Cheng, 2012) in the MBESS (Kelley, 2007b, 2007a, 2016) R (R Development Core Team, 2015) package. We used 1,000 bootstrap replications for constructing the bootstrap confidence intervals.[14] Our primary outcome variable was the proportion of the computed confidence intervals

[14] We used the OpenMx (Boker et al., 2011) package in R within the MBESS `ci.reliability()` function to compute the likelihood-based confidence intervals for the coefficient alpha and coefficient omega. For the ADF method, we wrote a function in R to analyze the ADF confidence interval for coefficient alpha using the equation given in the appendix of Maydeu-Olivares et al. (2007). For the ADF confidence interval for coefficient omega, we used the lavaan (Rosseel, 2012) package with the ADF (also termed weighted least square; WLS) as the method of estimation to run a one-factor CFA and used the model constraint command to estimate the confidence interval for coefficient omega.

Table 1
Types of Confidence Intervals Examined in Each Simulation Study for the Type of Reliability Coefficient

Type of confidence intervals	Coefficient alpha	Coefficient omega	Hierarchical omega	Categorical omega
Feldt	1	1		
Delta				
Normal-theory maximum likelihood				
Untransformed	1	1		
Logistic transformation	1	1		
Asymptotic distribution free				
Untransformed	1	1		
Logistic transformation	1	1		
Robust maximum likelihood				
Untransformed		1		
Logistic transformation		1		
Transformation				
Fisher	1	1		
Bonett	1	1		
Hakstian & Whalen	1	1		
Likelihood	1	1		
Bootstrap				
Bootstrap standard error				
Untransformed	1	1 & 2	2 & 3	3
Logistic transformation	1	1 & 2	2 & 3	3
Percentile	1	1 & 2	2 & 3	3
Bias corrected and accelerated	1	1 & 2	2 & 3	3

Note. The tabled numbers represent the simulation study or studies in which the particular reliability coefficient—in combination with the confidence interval method—is investigated.

that correctly bracketed the population parameter (i.e., empirical confidence interval coverage). For 95% confidence intervals, the average population coverage (i.e., the proportion of the computed confidence intervals that correctly bracketed the population parameter) should be close to .95. We will highlight the methods and conditions under which proportion coverage was within .925–.975 and .94–.96 ranges and labeled as "acceptable" and "good" coverage, respectively. The range of .925–.975 is equivalent to the Bradley (1978) liberal criterion ($1 - \alpha' \pm .5\alpha'$) which was used in Romano et al. (2010) and Padilla et al. (2012), among others. The proportion of acceptable and good coverage, however, does not account for the continuity of the coverage rate (e.g., the coverage .949 should be better than the value of .941).

To determine which of the design factors contributed to the coverage rate within each of the three studies, we used ANOVA where design factors are used as fixed factors. Similar to Lüdtke et al. (2008), ANOVA was conducted at the cell mean level where the average of a coverage rate of each cell is used as a dependent variable—one observation for each cell so the highest-level interaction (e.g., five-way interaction in Study 1) could not be separated from the error. We used the proportion of variance explained (η^2) as a measure of effect size for each of the main effects and interaction effects. We used two criteria to select the design factors that influenced the results: $\eta^2 \geq .01$ (Lüdtke et al., 2008) and $\eta^2 \geq .05$ (Geldhof, Preacher, & Zyphur, 2014). We found that the design factors with $.01 \leq \eta^2 < .05$ did not provide meaningful differences in the coverage rates shown later. Therefore, we considered the design factors with η^2 greater than or equal to .05 only.

All data were generated using the R environment for statistical computing and graphics (R Development Core Team, 2015). For the multivariate normal data, we used the `mvrnorm()` function from the MASS package (Venables & Ripley, 2015). Nonnormal data were generated with the `simulateData()` function from the lavaan package (Rosseel, 2012). We used 1,000 replications for each condition investigated.[15]

Study 1: Confidence Intervals in Congeneric Measurement Model Without Model Error at the Population Level

The objective of this study is to compare different confidence interval methods of coefficient alpha and coefficient omega in estimating population reliability. Hierarchical omega is not considered here because it is equal to coefficient omega in the population in perfectly fitting models. More specifically, the population value of coefficient omega and hierarchical omega are both equal to the population reliability. Categorical omega is not considered in Study 1 because we focus on continuous items in this study.

We conducted a Monte Carlo simulation study to compare the effectiveness of five classes of confidence interval methods that can further be divided into 13 methods for coefficient alpha and 15

[15] The 1,000 replications is large enough, such that the 95% confidence interval for the population of confidence interval coverage is sufficiently narrow. In particular, suppose that the observed proportion in a condition was 95% (i.e., the empirical coverage was equal to the nominal coverage). With 1,000 replications the 95% confidence interval for the population proportion would be 93.65% to 96.35% (based on the usual formula for a 95% two-sided confidence interval for a population proportion). Thus, the confidence interval for an observed proportion of .95 has a width of only .027 units (on the proportion scale; or 2.7 percentage points), which we regard as sufficiently small for evaluating the effectiveness of the various methods.

methods for coefficient omega. Thirteen methods for coefficient alpha included (a) Feldt's approach, (b) the normal-theory delta method, (c) the ADF delta method, (d) the normal-theory delta method with logistic transformation, (e) the ADF delta method with logistic transformation, (f) Fisher's *z* transformation, (g) Bonett's (2002) transformation, (h) Hakstian and Whalen (1976) transformation, (i) likelihood-based method, (j) the bootstrap standard error approach, (k) the bootstrap standard error approach with logistic transformation, and (l) percentile bootstrap, and (m) BC*a* bootstrap. All confidence interval methods for coefficient alpha are applicable for coefficient omega. Two additional methods for coefficient omega are (*n*) the normal-theory delta method estimated by robust maximum likelihood estimation and (o) the normal-theory delta method estimated by robust maximum likelihood estimation with logistic transformation. We use a confidence level of 95% because of its prevalence in the applied literature.

Method

The factors we examine here (in Study 1) are (a) sample size (five levels); (b) number of items (five levels); (c) factor loading distributions (two levels); (d) population reliability (three levels); and (e) item distributions (four levels). We fully crossed each of these five factors and thus a total of 600 distinct conditions were investigated. We briefly outline each of these five factors. There are 13 and 15 types of confidence interval procedures for coefficients alpha and omega, respectively, in Study 1.

Sample size. Sample size values of 50, 100, 200, 400, and 1,000 are used in the Monte Carlo simulation study. The maximum sample size of 1,000 is used to investigate what many researchers might consider a large-sample size. The values of 100, 200, and 400 are a compromise between the small sample size of 50 and the large sample size of 1,000.

Number of items. The number of items included on the homogeneous scales were 4, 8, 12, 16, and 20. The minimum number of items is 4 because confirmatory factor analysis will be underidentified with two items and will provide only perfectly fitting models in three items. The maximum number of items is 20, which is used for homogeneous scales in some contexts, such as ability tests.

Population reliability. The population reliability values we used are .7, .8, and .9. The population reliability of .7 has historically been considered an "acceptable value" of reliability in most areas in psychology. The reliability of .9 is generally considered to be a "high value" of reliability in much psychological research. Kline (2005) notes, although there is no "gold standard" to the interpretation of reliability, "reliability around .90 are considered 'excellent,' values around .80 are 'very good,' and values around .70 are 'adequate' (p. 50). We thus use values that we believe are reasonable values in psychological research.

Factor loading distribution. Factor loadings can be equal or unequal across items in a homogenous test. When factor loadings are equal across items, the scale is tau-equivalent and coefficient alpha estimates the population reliability. When factor loadings are not equal across items, the scale is not tau-equivalent and coefficient alpha and coefficient omega have different population values and coefficient alpha no longer estimates the population reliability. The performance of the two reliability indices, however, are not the same in the population with unequal factor loadings. In one condition the population factor loadings are equal and fixed to .5. To provide a range of values when factor loadings are not equal, the factor loadings increase from .2 to .8 with a step size of .6/($J-1$), where J is the number of items. The population error variances are calculated so that the population reliability is equal to the desired value (.7, .8, or .9) in the specified condition. Note that factor variance is always fixed to 1.

Item distribution. Data will be generated from four types of distribution, labeled D1–D4, which closely aligned with Enders (2001). D1 represents multivariate normality. D2–D4 are generated based on the Vale and Maurelli (1983) approach. All observed variables are set to have skewness of 1.25 and kurtosis of 3.5 in D2; skewness of 2.25 and kurtosis of 7 in D3; and skewness of 3.25 and kurtosis of 20 in D4. These can be nearly thought of as mild, moderate, and severe deviations from normality.

Results

This section will compare the effectiveness of the confidence interval methods for coefficient alpha and coefficient omega in bracketing the value of the population reliability. By "bracketing" we mean that the population value is contained within the lower and upper confidence interval limit (i.e., the population value is within the computed interval). There were some convergence issues in the Monte Carlo simulation study. The ADF method for omega estimation had convergence rates of less than 95% of the replications in 27% of the conditions; 24% of the conditions had no replications that converged. These problematic conditions tended to have sample sizes among the lower values studied (50–200) and for the high number of items of items studied (12–20). For other methods of omega estimation, less than 3% of the conditions had convergence rates of less than 95% of the replications. These conditions tended to have sample size among the lower value of those studied (50–100), among the lowest number of items studied (4–8), population reliability values of .7–.8, and with high levels of nonnormality. Therefore, the interpretation based on these results in these conditions should be done with an awareness of the nonconvergence issues. The table of convergence rates are available upon request from the authors.

All confidence interval methods of coefficient alpha had a poor performance in bracketing the value of the population reliability. The best method in this context was the bootstrap standard error method, which had only 50% the conditions that yielded acceptable coverage rates. In particular, coverage rates were poor when factor loadings were not equal and the number of items was low. Thus, we did not report the effects of design conditions on the confidence interval for coefficient alpha here (see the online supplement for the results) and recommended that the confidence interval for coefficient alpha should not be used in estimating population reliability.

Table 2 shows η^2 for all main and interaction effects of each design factor on the coverage rates of the confidence intervals for coefficient omega on population reliability. All three-way or higher-order interactions had η^2 lower than .05. The full table is available upon request.

Feldt's Approach. The η^2 of the main effects of the number of items and item distributions were greater than .05. The coverage rates were better when the number of items was higher: .854, .877, .888, .893, and .897 for 4, 8, 12, 16, and 20 items, respectively.

Table 2
The η^2 of the Effects of the Design Factors on the Coverage Rates of Coefficient Omega for Study 1

Factors	Feldt	NT	NT-L	ADF	ADF-L	NT-MLR	NT-MLR-L	Fisher	Bonett	HW	LL	BSE	BSE-L	PER	BC*a*
N	.001	.001	.002	**.077**	**.058**	**.166**	**.065**	.001	.002	.001	.003	.000	.010	**.123**	**.149**
J	**.056**	**.057**	.046	**.413**	**.429**	**.051**	.017	**.166**	**.051**	**.057**	.049	**.219**	**.131**	.031	**.093**
LOAD	.004	.003	.004	.000	.000	.001	.000	.004	.004	.004	.003	.003	.002	.000	.003
RELIA	.042	.033	.048	.003	.005	**.075**	**.189**	.001	.043	.041	.044	**.061**	**.167**	**.073**	.040
DIST	**.642**	**.652**	**.644**	.029	.037	**.290**	**.228**	**.540**	**.644**	**.641**	**.641**	**.159**	**.178**	**.183**	**.407**
N:J	.003	.003	.005	.038	.040	.021	.006	.005	.005	.003	.004	**.094**	**.050**	.016	.021
N:LOAD	.000	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000	.001	.001	.000	.000
N:RELIA	.002	.004	.001	.004	.005	.018	**.066**	.002	.002	.002	.001	.009	.046	.020	.004
N:DIST	.005	.007	.005	.003	.005	.006	.002	.002	.004	.004	.005	.000	.002	.000	.013
J:LOAD	.000	.000	.000	.000	.000	.000	.001	.001	.001	.000	.000	.000	.001	.000	.000
J:RELIA	.000	.000	.000	.001	.001	.005	.004	.000	.000	.000	.000	.033	.022	.000	.009
J:DIST	.044	.040	.040	.017	.019	.007	.002	**.075**	.042	.044	.042	.007	.001	.022	.003
LOAD:RELIA	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.001	.000
LOAD:DIST	.002	.002	.002	.000	.000	.000	.000	.003	.002	.002	.002	.000	.000	.000	.000
RELIA:DIST	.030	.029	.031	.005	.006	.028	.028	.010	.031	.030	.031	.014	.020	.011	.008

Note. The boldface numbers represent when $\eta^2 \geq .05$. All three-way and higher interactions are not tabled because their $\eta^2 < .05$. NT = normal-theory method; L = logistic-transformed Wald confidence interval; ADF = asymptotic distribution free; MLR = robust maximum likelihood estimation; HW = Hakstian-Whalen method; LL = likelihood-based method; BSE = bootstrap standard error; PER = percentile bootstrap; BC*a* = bias-corrected-and-accelerated bootstrap; *N* = sample size; J = number of items; LOAD = factor loading distribution; RELIA = population reliability; DIST = item distributions.

However, the convergence rates did not reach what we considered acceptable convergence levels (e.g., above 95%) regardless of the number of items. The coverage rates were worse when the item distributions were more deviated from normality: .945, .924, .827, and .831 for D1, D2, D3, and D4, respectively, that represented increasing levels of nonnormality (with D1 being multivariate normal). Thus, this method was not recommended, particularly for moderately to severely nonnormal item distributions.

Delta method. The delta method consists of three separate estimation methods: normal theory, ADF, and robust maximum likelihood, each using untransformed and logistic transformations.

Normal-theory (untransformed). The normal-theory approach was influenced by item distribution and the number of items. The coverage rates were .944, .922, .828, and .832 for D1, D2, D3, and D4, respectively, showing that there was a tendency for larger deviations from normality to worsen coverage. When the number of items was higher, the coverage rates were closer to .95 (i.e., .855, .876, .886, .893, and .897 for 4, 8, 12, 16, and 20 items, respectively). Thus, this method is not recommended for nonnormal item distributions.

Normal-theory with logistic transformation. This approach was influenced by item distributions such that the coverage rates were .946, .925, .829, and .833 for D1, D2, D3, and D4, respectively. Logistic transformation slightly improved the coverage rates.

ADF (untransformed). The main effects of sample size and the number of items were impactful on the coverage rates. More specifically, the coverage rates were better when sample size increased: .652, .654, .709, .773, and .825 for sample sizes of 50, 100, 200, 400, and 1000, respectfully. Further, the coverage rates were worse when the number of items increased: .909, .781, .673, .601, and .587 for 4, 8, 12, 16, and 20 items, respectfully. Although this method was not influenced much by the item distribution, the coverage rates were not acceptable in most cases.

ADF with logistic transformation. Similar to the standard ADF, the main effects of sample size and the number of items were impactful on the coverage rates. More specifically, the coverage rates were .681, .664, .712, .771, and .818 for sample sizes of 50, 100, 200, 400, and 1,000, respectively. Further, the coverage rates were .919, .786, .671, .599, and .587 for 4, 8, 12, 16, and 20 items, respectively. The logistic transformation did not always provide better coverage rates (e.g., .771 vs .773 for sample size of 400).

Robust maximum likelihood (untransformed). The main effects of sample size, the number of items, population reliability, and item distributions were impactful on the coverage rates. More specifically, the coverage rates tended to be slightly better when sample size increased: .923, .926, .924, .937, and .940 for sample sizes of 50, 100, 200, 400, and 1,000, respectively. Further, the coverage rates were .924, .928, .931, .933, and .934 when the numbers of items were 4, 8, 12, 16, and 20, respectively. The coverage rates decreased when population reliability was higher: .935, .930, and .924 for .7, .8, and .9, respectively. Lastly, the coverage rates were slightly lower when item distributions deviated more from normality: .941, .936, .922, and .921 for D1, D2, D3, and D4, respectively. Thus, robust maximum likelihood estimation helped the confidence intervals to provide coverage rates close to the acceptable range for nonnormal items and retained good coverage rates for normally distributed items. Because the main effects did not highly influence coverage rates (i.e., the largest difference was .02) and the coverage rates were close to the acceptable range in most cases, we find that robust maximum likelihood is preferred to standard maximum likelihood (normal theory).

A peculiarity in the above description of the results of the simulation study was that the coverage rates *decreased* when the population reliability was higher. We further investigated this issue and found that it is, at least in part, due to the asymmetry of the sampling distribution of the sample reliability coefficient, which is theoretically bounded at 1.0. For high values of reliability coefficients (which also occurs for Pearson product–moment correlation coefficients), the sampling distribution of sample estimates will be negatively skewed (the extent to which depends on sample size and the true population value; recall Footnote 8). Although the robust maximum likelihood approach does not assume that the items are multivariate normally distributed, it does assume that the sample reliability coefficient is normally distributed. Yet, as noted, the distribution of the sample reliability coefficient will be negatively skewed and thus when the population reliability is high sample values can be markedly lower than the population value (as compared with a normally distributed quantity). The confidence interval for the robust maximum likelihood estimator leads to a symmetric confidence interval whose upper bound does not extend as far as it ideally would and there are more "misses" of the population value on the high side for higher values of the population reliability coefficient. In particular, we find in our simulation that 3.5% of the time for the population reliability of .90 condition that the upper limit is *less* than the true value (i.e., high side misses). However, when the population reliability value is .70, only 2.4% of the intervals have upper limits less than the true value. For reliability of .80, we find that 2.9% of the confidence intervals have upper limits less than the true value (a value in between the other two, which makes sense given that .80 is between the other two population values). Separately, the lower limit of the confidence interval excludes the true value (a) more than is ideal but (b) essentially the same amount: 4.1%, 4.1%, and 4.0%, respectively, for .70, .80, and .90 population reliability values. Thus, the .935 coverage rate for the lower reliability condition is due to the proportion of lower confidence interval limits that are larger than the true value (i.e., .041) and upper confidence interval limits that are smaller than the true value (i.e., .024). Although the lower confidence interval cover is not ideal (in that they are ideally .025), the upper confidence interval limit misses more as population reliability is higher.

Robust maximum likelihood with logistic transformation. The main effect of item distribution and the interaction effect of sample size and population reliability were impactful on the coverage rates. Table 3 shows the coverage rates illustrating the interaction between sample size and population reliability. When sample size was low and population reliability was high, the coverage rates were not acceptable. Item distributions slightly influenced coverage rates: .942, .939, .925, and .926 for D1, D2, D3, and D4, respectively. However, note that the coverage was acceptable across the four distributions studied. The coverage rates were slightly better when logistic transformation was used.

Table 3
The Coverage Rates of Confidence Intervals for Coefficient Omega Using the Normal-Theory Method With Robust Maximum Likelihood With Logistic Transformation Classified by Sample Size and Population Reliability for Study 1

	Population reliability		
Sample size	.7	.8	.9
50	.953	.936	.915
100	.942	.926	.917
200	.929	.924	.921
400	.941	.938	.934
1,000	.941	.940	.938

Transformation-based approaches. The transformation based approaches consist of three separate transformations: Fisher, Bonett, and Hakstian, and Whalen.

Fisher's transformation. The interaction between the number of items and item distributions was impactful. Specifically, as shown in Table 4, the coverage rates were lower when items deviated from normality and higher when the number of items increased. However, the coverage rates in most conditions were higher than .975 or lower than .925. Because of the inconsistency of its performance, the Fisher's method is not recommended.

Bonett's transformation. The main effects of the number of items and item distributions were impactful. Specifically, when the number of items increased, the coverage rates were better: .856, .879, .888, .894, and .897 for 4, 8, 12, 16, and 20 items, respectfully. None of which, however, were in the acceptable range. When items were not normally distributed, the coverage rates were worse (i.e., .946, .925, .828, and .831 for D1, D2, D3, and D4, respectively). Thus, the Bonett's transformation was not robust to nonnormality.

Hakstian and Whalen's transformation. Similar to Bonett's transformation, the main effects of the number of items and item distributions had η^2 greater than .05 and were thus impactful. Specifically, the coverage rates were .854, .877, .887, .894, and .897 for 4, 8, 12, 16, and 20 items, respectively. Although the coverage rates improved for more items, none of which were acceptable. The coverage rates were .945, .924, .827, and .831 for D1, D2, D3, and D4, respectively. Thus, Hakstian and Whalen's transformation was not robust to nonnormality.

Likelihood-based approach. The main effect of item distributions was impactful. Specifically, when items deviated more from normality, the coverage rates tended to decrease: .946, .925, .829, and .834 for D1, D2, D3, and D4, respectively. The likelihood-based approach was thus not robust to nonnormality.

Bootstrap approach. The bootstrap approach consists of three separate types: bootstrap standard error, percentile, and BC*a*.

Bootstrap standard error (untransformed). The main effects of population reliability and item distribution and the interaction effect between sample size and the number of items had η^2 higher than .05. The coverage rates were slightly decreased when population reliability was higher (i.e., .944, .940, and .935 for .7, .8, and .9, respectively). The coverage rates were .947, .943, .936, and .932 for D1, D2, D3, and D4, respectively. Table 5 shows the interaction between the number of items and sample size. Most conditions had acceptable ranges of coverage rates except one with four items and a sample size of 50 or 100. Because the effects of design factors were not large, the bootstrap standard error was recommended for both normal and nonnormal data.

Bootstrap standard error with logistic transformation. Similar to the bootstrap standard error (directly above), the main effects of population reliability and item distribution and the interaction effect between sample size and the number of items had η^2 higher than .05. The logistic transformation provided better coverage rates than the bootstrap standard error without logistic transformation (e.g., see Table 5 of the online supplement). However, coverage, rates were not acceptable for the equal factor loading conditions with four items, and for the unequal factor loading conditions rates were not acceptable for 4, 8, or 12 items (yet they were acceptable for 16 and 20 items).

Table 4
The Coverage Rates of Confidence Intervals for Coefficient Omega Using Fisher's Method Classified by the Number of Items and Item Distributions for Study 1

	Number of items				
Item distributions	4	8	12	16	20
D1	.965	.980	.980	.982	.981
D2	.939	.966	.972	.975	.977
D3	.848	.889	.904	.912	.916
D4	.805	.884	.912	.926	.935

Note. D1–D4 in the item distributions column represents the particular distributional form. D1 is multivariate normal. D2 has a skewness of 1.25 and kurtosis of 3.5. D3 has a skewness of 2.25 and kurtosis of 7. D4 has a skewness of 3.25 and kurtosis of 20.

Percentile bootstrap. The main effects of population reliability and item distributions had η^2 greater than .05. When the population reliability was higher, the coverage rates were slightly lower (.938, .933, and .930 for population reliability of .7, .8, and .9, respectively). The coverage rates were .942, .938, .926, and .930 for D1, D2, D3, and D4, respectively. In general, percentile bootstrap had an acceptable coverage rate in both normal and nonnormal item distributions.

BCa bootstrap. The main effects of sample size, the number of items, and item distributions were impactful. Specifically, when sample size increased, the coverage rates became (slightly) better: .918, .923, .922, .933, and .937 for sample sizes of 50, 100, 200, 400, and 1,000, respectively. When the number of items increased, the coverage rates were also slightly better (.918, .924, .927, .931, and .933 for 4, 8, 12, 16, and 20 items). The coverage rates were .942, .925, .920, and .913 for D1, D2, D3, and D4, respectively. Thus, the BC*a* method had an acceptable coverage rate for most conditions with both normal and nonnormal items; however, the coverage rates were not as good as the bootstrap standard error and percentile bootstrap.

In conclusion, all methods except ADF and Fisher's approaches provided acceptable coverage rates for normal items. However, only normal-theory approach with robust standard error and all bootstrap methods provided acceptable coverage rates for most conditions of nonnormal items. Logistic transformation slightly improved the coverage rates for the normal-theory approach with robust standard error and bootstrap standard error.

Study 2: Confidence Intervals for Congeneric Measurement Model With Small Model Error at the Population Level

In the congeneric measurement model with small model error, not coefficient omega but rather hierarchical omega represents population reliability. In this study, the performance of confidence intervals for coefficient omega and hierarchical omega is investigated in terms of whether they appropriately bracket the population reliability. From the previous study, bootstrap methods performed well in both normal and nonnormal items so we evaluate bootstrap methods in this study. The normal-theory approach was not used in this study because it is not available for hierarchical omega, as hierarchical omega cannot be derived from the CFA model parameters (it requires the observed covariance matrix of the items or the observed variance of *Y*). That is, only bootstrap confidence intervals for coefficient omega and hierarchical omega, including (a) the bootstrap standard error approach, (b) the bootstrap standard error approach with logistic transformation, (c) percentile bootstrap, and (d) BC*a* bootstrap, are compared in this study. We use a confidence level of 95% because of its prevalence in the applied literature.

Method

We designed the simulation of Study 2 similarly to Study 1 but with four design factors used: sample size, number of items, population coefficient omega, and model error. However, here we only investigate bootstrap methods.

Sample size. Sample sizes used (*N*) are 50, 100, 200, 400, and 1,000.

Number of items. The number of items (*J*) included on the homogeneous scales were 4, 8, and 12. We removed the 16 and 20 items because three levels of the number of items were sufficient to investigate the influence of the number of items. The coverage rate differences between 12, 16, and 20 items in the previous study were not large in bootstrap confidence intervals.

Population coefficient omega. The population coefficient omega includes .7, .8, and .9. We used the unequal factor loading specified in Study 1 such that the factor loadings increase from .2 to .8 with a step size of.6/$(J - 1)$ (we did not include the equal loading distribution condition because its influence on the coverage rates of confidence intervals for coefficient omega was not high in Study 1). Then, measurement error variances of each item were calculated such that the coefficient omega was equal to the specified condition. The factor variance was always fixed to 1.

Model error. The amount of model error is quantified by the population root mean square error of approximation (RMSEA). The values of the population RMSEA used are .02, .05, .08, and .10. The population RMSEA (ε) is calculated by

$$\varepsilon = \sqrt{F_0/df}, \tag{38}$$

with F_0 defined as

$$F_0 = \mathrm{tr}(\boldsymbol{\Sigma}[\boldsymbol{\Sigma}_{\mathrm{M}}]^{-1}) - \log|\boldsymbol{\Sigma}[\boldsymbol{\Sigma}_{\mathrm{M}}]^{-1}| - J, \tag{39}$$

where $\boldsymbol{\Sigma}$ is the observed population covariance matrix, $\boldsymbol{\Sigma}_{\mathrm{M}}$ is the model-implied covariance matrix, and *df* is the model's degree of freedom. To find $\boldsymbol{\Sigma}$ that yields a specified RMSEA, we do the following multiple step procedure: (a) Let $\mathbf{R}_{\mathrm{E}}$ be the $J \times J$

Table 5
The Coverage Rates of Confidence Intervals for Coefficient Omega Using the Bootstrap Standard Error Method (Untransformed) Classified by the Number of Items and Sample Size for Study 1

	Number of items				
Sample size	4	8	12	16	20
50	.929	.943	.953	.961	.967
100	.928	.937	.945	.951	.955
200	.932	.932	.934	.937	.939
400	.939	.939	.939	.940	.944
1,000	.938	.942	.941	.943	.941

correlation matrix of errors and $\mathbf{V}_{\mathrm{E}}$ be a $J \times J$ diagonal matrix of measurement error variances. The goal is to find $\mathbf{R}_{\mathrm{E}}$ such that $\boldsymbol{\Sigma}$ provides the specified RMSEA. (b) $\boldsymbol{\Sigma}$ is calculated by $\boldsymbol{\Lambda\Lambda}' + \mathbf{V}_{\mathrm{E}}^{1/2}\mathbf{R}_{\mathrm{E}}\mathbf{V}_{\mathrm{E}}^{1/2}$. (c) $\boldsymbol{\Sigma}$ is fitted using the congeneric measurement model so $\boldsymbol{\Sigma}_{\mathrm{M}}$ is obtained. (d) ε is calculated from Equation 38. A numerical method is used to solve for $\mathbf{R}_{\mathrm{E}}$, which provides the specified population RMSEA.[16] Note that there are many sets of $\mathbf{R}_{\mathrm{E}}$ that provide the same population RMSEA. We picked one solution by using a random number seed.

In conclusion, there are a total of (5 × 3 × 3 × 5) 225 distinct conditions for each of the four types of confidence interval procedures for coefficients omega and hierarchical omega. For each of the 225 conditions, 1,000 replications were used. Note that population reliability is equal to hierarchical omega, which is not exactly equal to, although close to, coefficient omega specified in the previous design factors. We check whether each confidence interval is bracketing the value of the population reliability (i.e., hierarchical omega) and not the population coefficient omega.

Results

This section will compare the effectiveness of the confidence interval methods for coefficient omega and hierarchical omega in bracketing the value of the population reliability. Based on each of the confidence interval methods evaluated in the 225 conditions, 15 conditions had convergence rates less than 95%. These conditions tend to have low sample size (50–100) and 4 items. The table of convergence rates is available upon request from the authors.

All confidence interval methods of coefficient omega had poor performances in bracketing the value of the population reliability. Only 52%, 71%, 72%, and 64% of all conditions from the confidence intervals for coefficient omega using bootstrap standard error, bootstrap standard error with logistic transformation, percentile bootstrap, and BC*a* bootstrap had acceptable coverage rates (compared with 93%, 92%, 99%, and 96% in confidence intervals for hierarchical omega). In particular, the differences between the coverage rates of coefficient omega and hierarchical omega were the largest when sample size was small and the population coefficient omega was .9. Thus, we did not report the effects of design conditions on the confidence intervals for coefficient omega here (see the online supplement for those results) and recommend that the confidence intervals for coefficient omega not be used in estimating population reliability when a CFA model does not perfectly fit data.

Table 6 shows η^2 for all main and interaction effects of each design condition on the coverage rates of the confidence intervals for hierarchical omega on population reliability.

Bootstrap standard error. The interactions between (a) sample size and the number of items and (b) population reliability and the number of items had η^2 values greater than .05. As shown in Table 7, when sample size is small, the coverage rates were lower than .95 for four items but higher than .95 for 12 items. The coverage rates were closer to .95 when sample size increased. As also shown in Table 7, when population reliability is .7, the coverage rates were lower than .95 in the four-item conditions but higher than .95 in the 12-item conditions. The coverage rates were closer to .95 when population reliability was .9. All conditions had coverage rates in the acceptable range.

Table 6
The η^2 *of the Effects of the Design Factors on the Coverage Rates of Hierarchical Omega for Study 2*

Factors	BSE	BSE-L	PER	BC*a*
N	.004	**.070**	**.051**	**.152**
J	**.383**	**.262**	.000	**.187**
RELIA	.008	.041	**.102**	.003
RMSEA	.018	.011	.006	.002
N:J	**.090**	**.090**	.008	**.051**
N:RELIA	.006	.009	.042	.000
N:RMSEA	.001	.000	.000	.003
J:RELIA	**.062**	.044	.001	**.064**
J:RMSEA	.001	.001	.007	.002
RELIA:RMSEA	.001	.000	.000	.001

Note. The boldface numbers represent the $\eta^2 \geq .05$. All interactions higher than two way are not presented here because their $\eta^2 < .05$. L = logistic-transformed Wald confidence interval; BSE = bootstrap standard error; PER = percentile bootstrap; BC*a* = bias-corrected-and-accelerated bootstrap; *N* = sample size; J = number of items; RELIA = population coefficient omega; RMSEA = root mean square error of approximation.

Bootstrap standard error with logistic transformation. The interaction between sample size and the number of items had η^2 greater than .05. The pattern is similar to one in bootstrap standard error without logistic transformation (directly above): interactions between (a) sample size and the number of items and (b) population reliability and the number of items had η^2 values greater than .05. The transformation made the coverage rates closer to .95, as compared with the untransformed approach, in most conditions. However, in the sample size of 50 and 12 items, the coverage rate was farther from .95 and was not in the acceptable range. Therefore, logistic transformation is recommended except in small sample size conditions and large number of items.

Percentile bootstrap. The main effects of sample size and population reliability were impactful. Specifically, when the coverage rates were .947, .942, .939, .948, and .948 for sample sizes of 50, 100, 200, 400, and 1,000, respectively. Further, the coverage rates were .948, .945, and .942 when population reliability was .7, .8, and .9, respectively. All conditions had coverage rates in the acceptable range and most of them had coverage rates in the "good" range.

BC*a* bootstrap. The interaction effects between (a) sample size and the number of items and (b) population reliability and the number of items were greater than .05. As shown in Table 8, when sample size is low, the coverage rates were closer to .95 as the number of items increased. However, the number of items did not affect coverage rates in large sample size. For four items, the coverage rates increased when population reliability increased. For 12 items, however, the coverage rates decreased when population

[16] Note that our method of simulating data for misspecified models is similar to Cudeck and Browne (1992) but there is a distinction. Consider a population covariance matrix defined as $\boldsymbol{\Sigma} = \mathbf{C} + \mathbf{E}$, where $\mathbf{C}$ is the population covariance matrix without model errors and $\mathbf{E}$ is the population model error. Let $\boldsymbol{\Sigma}_{\mathrm{M}}$ be the model-implied covariance matrix after fitting $\boldsymbol{\Sigma}$ to the model. Cudeck and Browne (1992) use a numerical method to find the model error that $f(\boldsymbol{\Sigma}_{\mathrm{M}}, \mathbf{C}) = c$, where c is a constant. This is not the same as the definition of RMSEA based on $f(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_{\mathrm{M}})$. Our method uses a numerical method to model error such that $f(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_{\mathrm{M}}) = c$, which is consistent with the definition of RMSEA.

Table 7

The Coverage Rates of Confidence Intervals for Hierarchical Omega Using the Bootstrap Standard Error Method Classified by (a) the Number of Items and Sample Size (Top Section) and (b) the Number of Items and Population Coefficient Omega (Bottom Section) for Study 2

		Number of items		
Factors	Levels	4	8	12
Sample size	50	.931	.958	.972
	100	.933	.956	.963
	200	.945	.943	.948
	400	.948	.953	.957
	1,000	.944	.951	.951
Population coefficient omega	.7	.937	.954	.964
	.8	.940	.952	.958
	.9	.944	.950	.953

reliability increased. Most conditions had coverage rates in the acceptable range.

In conclusion, all of the methods we studied for forming confidence intervals for hierarchical omega tended to have good coverage rates for bracketing the population reliability value. The different confidence interval methods tended to produce rather trivial differences in coverage rates for hierarchical omega. Nevertheless, the percentile bootstrap method had more instances of good coverage and no issues of poor performance for the conditions investigated. Thus, we recommend the use of the percentile bootstrap method when forming confidence intervals for the population reliability coefficient when using hierarchical omega in the context of misspecified models. Additionally, we remind readers that the performance of coefficient omega when the model is misspecified produced poor results and thus coefficient omega is not recommend for use with misspecified models.

Study 3: Confidence Intervals for Congeneric Measurement Model With Categorical Items

When items are ordered categorical, the relationship between factor and observed item scores is not linear. Rather, the factor is thought to have a linear relationship with continuous latent variables underlying each item, which are categorized via thresholds to yield scores for each category. Instead of coefficient omega or hierarchical omega, categorical omega is appropriate to represent scale reliability for ordered categorical items. This simulation study will investigate the performance of confidence intervals for hierarchical omega and categorical omega in bracketing the value of the population reliability for categorical items. Coefficient omega is not considered here because we assume that all models have some degree of model misspecification, correspondingly, model error is added in this simulation. Bootstrap confidence intervals are tested in this simulation because they are available for both hierarchical and categorical omega. That is, four methods are compared in this study: (a) the bootstrap standard error, (b) the bootstrap standard error with logistic transformation, (c) percentile bootstrap, and (d) BC*a* bootstrap. We use a confidence level of 95% because of its prevalence in the applied literature.

Factors

We designed the simulation conditions to be similar to Study 1 and 2. Five design factors were used in this study: sample size, number of items, number of categories, threshold symmetry, and population categorical omega (for perfectly fitting model). However, we only investigate hierarchical omega and categorical omega.

Sample size. The sample sizes (*N*) included 50, 100, 200, 400, and 1,000.

Number of items. The number of items (*J*) included on the homogeneous scales were 4, 8, and 12.

The number of categories. The numbers of categories considered are two or five, both of which are commonly used in applied research.

Threshold symmetry. We used five levels of threshold symmetry following Rhemtulla, Brosseau-Liard, and Savalei (2012): symmetry, moderate asymmetry, extreme asymmetry, moderate asymmetry-alternating, and extreme asymmetry-alternating. In the symmetry conditions, thresholds are distributed evenly around 0 and spaced evenly to divide the distance between −2.5 and 2.5. In the moderate asymmetry conditions, the thresholds are created such that the peak category is on the left near the center. In the extreme asymmetry conditions, the thresholds are created such that the peak category is in the lowest category. See Figure 2 for the distributions of each category when different threshold symmetries are imposed. The alternating conditions are that the odd-number items had the reversed distributions.

Population categorical omega for perfectly fitting model. The population categorical omega values included .7, .8, and .9. Similar to the previous studies, the factor variance is always fixed to 1 and the factor loadings are unequal using the same values as in Study 1 and 2. Then, the error variances of each latent variable underlying each item are calculated such that the population categorical omega value is equal to the value noted in the condition. Note that the scale of the model-implied covariance matrix of the hypothesized latent variable underlying each item is transformed to have a total variance of 1 before calculating categorical omega following delta parameterization. Thus, the rescaled factor loadings and measurement error variances that provide total variances of 1 are used for data generation.

Table 8

The Coverage Rates of Confidence Intervals for Hierarchical Omega Using the Bias-Corrected-and-Accelerated Bootstrap Classified by (a) the Number of Items and Sample Size (Top Section) and (b) the Number of Items and Population Coefficient Omega (Bottom Section) for Study 2

		Number of items		
Factors	Levels	4	8	12
Sample size	50	.924	.938	.950
	100	.936	.943	.948
	200	.940	.939	.941
	400	.944	.947	.949
	1,000	.946	.951	.949
Population coefficient omega	.7	.935	.945	.951
	.8	.939	.944	.948
	.9	.941	.941	.944

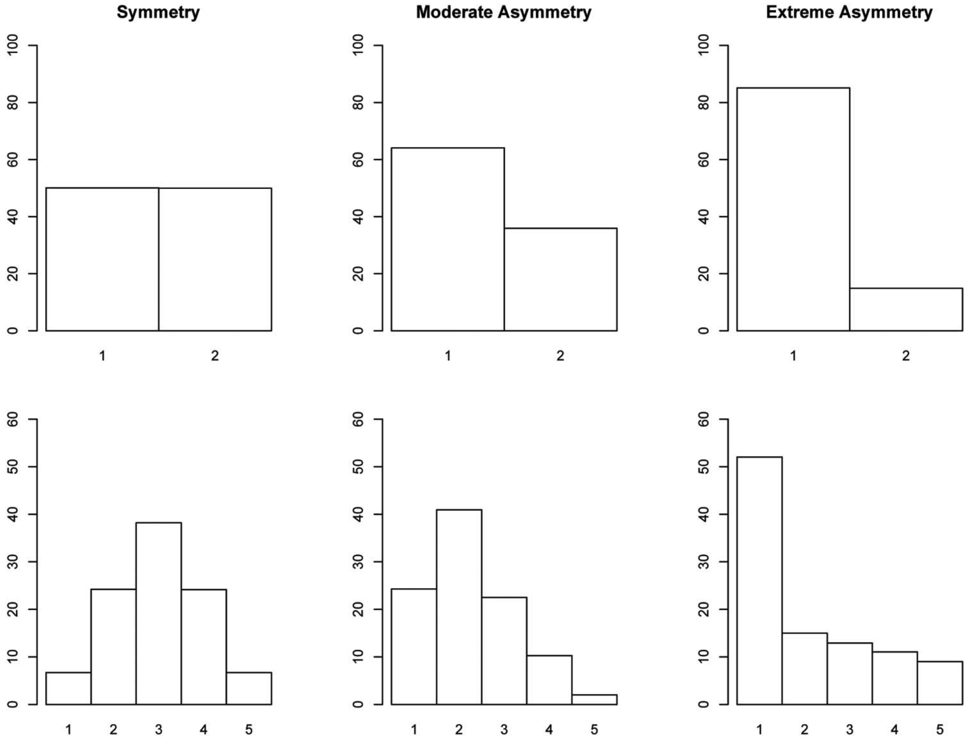

Figure 2. The item distributions of each level of threshold pattern (symmetry, moderate asymmetry, or extreme asymmetry) and the number of categories (two or five).

Then, measurement error correlations of latent variables underlying categorical items are added such that the population RMSEA is .05. Thus, the actual categorical omega is not exactly equal but close to .7, .8, or .9. In conclusion, there are a total of (5 × 3 × 2 × 5 × 3) 450 distinct conditions for each of the four types of confidence interval procedures for hierarchical omega and categorical omega. In each of the 600 conditions, 1,000 replications were used.

Results

This section will compare the effectiveness of the confidence interval methods for hierarchical omega and categorical omega in bracketing the value of the population reliability. Based on each of the confidence interval methods evaluated in the 450 conditions, 193 conditions of hierarchical omega (43%) and 168 conditions of categorical omega (37%) had convergence rates less than 95%. We did not find systematic patterns in the proportion of nonconvergent results across design conditions. The most influential design condition was sample size. Sample size slightly decreased the proportion of nonconvergent results in both types of confidence intervals. When sample sizes were 50, 100, 200, 400, and 1,000, the proportion of nonconvergence results were 15%, 12%, 11%, 11%, and 11% for hierarchical omega and 13%, 12%, 11%, 11%, and 11% for categorical omega. The table of convergence rates is available upon request from the authors.

The performances of the confidence interval methods for hierarchical omega were slightly worse than the performance of confidence interval methods for categorical omega with regards to bracketing the value of the population reliability. That is, 68%, 62%, 57%, and 53% of all conditions from the confidence intervals for hierarchical omega using bootstrap standard error, bootstrap standard error with logistic transformation, percentile bootstrap, and BC*a* bootstrap bracketed the population reliability, compared with 40%, 51%, 36%, and 74% from the confidence intervals for categorical omega. We will provide the results for both types of confidence intervals.

Table 9 shows η^2 for all main and interaction effects of each design factors on the coverage rates of the confidence intervals for hierarchical omega and categorical omega on population reliability.

Confidence intervals for hierarchical omega.

Bootstrap standard error. Three two-way interaction effects had η^2 higher than .05. These two-way interaction effects involved with threshold patterns with other design factors: (a) population categorical omega for perfectly fitting model (see Table 10); (b) the number of categories (see Table 11); and (c) the number of items (see Table 12. In threshold patterns 4 and 5, the coverage rates were low and out of the acceptable range when population reliability was high or items were dichotomous or the number of items was low. Therefore, the bootstrap standard error confidence interval for hierarchical omega is not recommended for scales containing items with different threshold patterns.

Bootstrap standard error with logistic transformation. The results were similar to bootstrap standard error (see directly above). Logistic transformation did not improve the coverage rate so this method is not recommended.

Table 9
The η^2 of the Effects of the Design Factors on the Coverage Rates of Hierarchical Omega and Categorical Omega for Study 3

	Hierarchical omega				Categorical omega			
Factors	BSE	BSE-L	PER	BC*a*	BSE	BSE-L	PER	BC*a*
N	.021	.014	.003	.004	.001	**.065**	.019	.015
J	**.050**	**.050**	.022	.047	**.093**	.028	**.154**	.020
RELIA	.048	**.065**	**.072**	**.066**	**.162**	**.088**	**.203**	.048
NCAT	.037	.038	**.052**	.040	.010	.007	**.077**	**.055**
THRES	**.280**	**.298**	**.336**	**.324**	.013	.006	.008	.020
N:J	.001	.001	.004	.001	.003	.016	.004	.012
N:RELIA	.001	.000	.001	.000	.004	.016	.007	.000
N:NCAT	.001	.001	.000	.001	.000	.000	.008	.011
N:THRES	.022	.018	.010	.014	.000	.000	.005	.000
J:RELIA	.003	.002	.002	.002	**.102**	**.073**	**.119**	**.063**
J:NCAT	.002	.001	.000	.001	.005	.005	.017	.000
J:THRES	**.054**	.049	.032	.045	.008	.034	.007	.021
RELIA:NCAT	.001	.001	.001	.001	.000	.000	.021	.016
RELIA:THRES	**.063**	**.063**	**.055**	**.062**	.014	.015	.010	.005
NCAT:THRES	**.061**	**.061**	**.074**	**.061**	.014	.003	.023	.030

Note. The boldface numbers represent when $\eta^2 \geq .05$. All interactions higher than two ways are not presented here because their $\eta^2 < .05$. L = logistic-transformed Wald confidence interval; BSE = bootstrap standard error; PER = percentile bootstrap; BC*a* = bias-corrected-and-accelerated bootstrap; *N* = sample size; J = number of items; RELIA = population categorical omega for perfectly fitting model; NCAT = number of categories; THRES = threshold pattern.

Percentile bootstrap. The interactions between (a) threshold patterns and population categorical omega for perfectly fitting model, and (b) threshold patterns and the number of categories had η^2 higher than.05. As shown in Table 10 and 11, in threshold patterns 4 and 5, the coverage rates were low and out of acceptable range when population reliability was high or items were dichotomous. This method is not recommended for scale containing items with different threshold patterns.

Table 10
The Coverage Rates of Confidence Intervals For Hierarchical Omega Using Four Bootstrap Methods Classified by Population Coefficient Omega and Threshold Patterns for Study 3

		Threshold patterns				
Methods	Population categorical omega for perfectly fitting models	1	2	3	4	5
BSE	.7	.960	.957	.955	.943	.725
	.8	.958	.956	.957	.890	.543
	.9	.954	.948	.965	.694	.223
BSE-L	.7	.962	.958	.956	.942	.709
	.8	.953	.952	.953	.874	.505
	.9	.930	.927	.946	.645	.185
PER	.7	.951	.947	.943	.932	.650
	.8	.940	.940	.936	.857	.403
	.9	.902	.900	.910	.584	.147
BC*a*	.7	.944	.941	.940	.922	.673
	.8	.939	.939	.940	.852	.455
	.9	.910	.912	.926	.610	.161

Note. L = Logistic-transformed Wald confidence interval; BSE = Bootstrap standard error; PER = Percentile bootstrap; BC*a* = Bias-corrected-and-accelerated bootstrap. The threshold patterns 1–5 represent, symmetry, moderate asymmetry, extreme asymmetry, moderate asymmetry-alternating, and extreme asymmetry-alternating, respectively. For the "asymmetry-alternating" conditions the odd-numbered items had reversed distributions (i.e., the skew alternated). See Figure 2 for visual representations of the threshold patterns.

BCa bootstrap. The results were similar to percentile bootstrap so we do not repeat it again.

Confidence intervals for categorical omega.

Bootstrap standard error. The interaction between population reliability and the number of items had η^2 greater than .05. As shown in Table 13, when population reliability was .7, the coverage rates were relatively constant across the number of items. However, with the population reliability of .9, the coverage rates were lower when the number of items increased. The coverage rates were not acceptable when population reliability and the number of items were both high.

Table 11
The Coverage Rates of Confidence Intervals for Hierarchical Omega Using Four Bootstrap Methods Classified by the Number of Categories and Threshold Patterns for Study 3

		Threshold patterns				
Methods	Number of categories	1	2	3	4	5
BSE	2	.959	.957	.963	.777	.297
	5	.956	.951	.955	.914	.699
BSE-L	2	.948	.947	.956	.748	.262
	5	.948	.943	.947	.898	.672
PER	2	.928	.927	.927	.708	.161
	5	.933	.931	.932	.881	.640
BC*a*	2	.930	.931	.940	.722	.223
	5	.932	.930	.931	.874	.639

Note. L = logistic-transformed Wald confidence interval; BSE = bootstrap standard error; PER = percentile bootstrap; BC*a* = bias-corrected-and-accelerated bootstrap. The threshold patterns 1–5 represent, symmetry, moderate asymmetry, extreme asymmetry, moderate asymmetry-alternating, and extreme asymmetry-alternating, respectively. For the "asymmetry-alternating" conditions the odd-numbered items had reversed distributions (i.e., the skew alternated). See Figure 2 for visual representations of the threshold patterns.

Table 12
The Coverage Rates of Confidence Intervals for Hierarchical Omega Using Bootstrap Standard Error Classified by the Number of Items and Threshold Patterns for Study 3

	Threshold patterns				
Number of items	1	2	3	4	5
4	.947	.943	.954	.689	.267
8	.961	.956	.961	.898	.505
12	.964	.962	.963	.949	.731

Note. The threshold patterns 1–5 represent, symmetry, moderate asymmetry, extreme asymmetry, moderate asymmetry-alternating, and extreme asymmetry-alternating, respectively. For the "asymmetry-alternating" conditions the odd-numbered items had reversed distributions (i.e., the skew alternated). See Figure 2 for visual representations of the threshold patterns.

Bootstrap standard error with logistic transformation. The interaction effect between population reliability and the number of items and the main effect of sample size had η^2 greater than .05. For the interaction, as shown in Table 13, the pattern is similar to the one in bootstrap standard error without logistic transformation. When sample size increased, the coverage rates were lower: .967, .930, .916, .918, and .914 for 50, 100, 200, 400, and 1,000, respectively. The correction made the coverage rates closer to .95 in most conditions. However, the coverage rates were still not acceptable when both population reliability and the number of items were high.

Percentile bootstrap. The interaction effect between the number of items and population reliability was impactful. Specifically, as shown in Table 13, the pattern of coverage was similar to that in bootstrap standard error; however, the coverage rates were worse than when using bootstrap standard error, especially in the conditions in which the number of items and population reliability were both high. The main effect of the number of categories was also impactful, such that the coverage rate was better with five categories (.891) than in two categories (.810), although, coverage was not at acceptable levels in either case.

BCa bootstrap. The interaction effect between the number of items and population reliability and the main effect of the number of categories had η^2 greater than .05. As shown in Table 13, when the number of items increased, the coverage rates increased for population reliability of .7 and for population reliability of .8, but decreased for population reliability of .9. The coverage rate for five categories (.954) was better than the coverage rate for two categories (.933). Most conditions had coverage rates in the acceptable range except when both the number of items and population reliability were high. The coverage rate in this condition was still better than it would have been using other methods.

In conclusion, the BC*a* confidence interval method for categorical omega is recommended for estimating population reliability for scales with categorical items. Confidence intervals for hierarchical omega are appropriate in conditions in which the threshold patterns are the same across items. In our experience these conditions are rare in practice.

Discussion

In this article we discussed four different reliability coefficients for quantifying the reliability of a composite score. We then acknowledge that the population reliability coefficient is what is ultimately desired, not the sample value, and that a confidence interval should accompany the estimated reliability coefficient. However, there are many methods of confidence interval formation that can accompany a variety of reliability coefficients. We evaluate the various confidence interval methods in a variety of conditions with three Monte Carlo simulation studies in order to make recommendations to researchers about which confidence interval methods perform most effectively.

Researchers frequently use coefficient alpha to quantify the reliability of composite scores. However, when coefficient alpha is used to estimate the population reliability coefficient alpha assumes, in addition to the standard classical test theory assumptions, that (a) the factor loadings of each item are equal (i.e., tau-equivalence holds); (b) one-factor CFA model must perfectly fit the item covariances; and (c) items are continuous. If these conditions do not hold in the population, the population value for coefficient alpha is not the population value of reliability.

Many methodologists have promoted coefficient omega, which relaxes the assumption of equal factor loadings (e.g., Kelley & Cheng, 2012; Marcoulides & Saunders, 2006; Padilla & Divers, 2013a; Raykov, 1997; Raykov & Shrout, 2002). In this article we go further than just recommending coefficient omega. In particular, we encourage researchers to use hierarchical omega for continuous items. Hierarchical omega does not require a perfectly fitting CFA model. As MacCallum and Austin (2000) stated, "all models are wrong to some degree, even in the population, and that the best one can hope for is to identify a parsimonious substantively meaningful model that fits observed data adequately well" (p. 218). That is, a perfect homogeneous model is not reasonable to assume for a set of items, even in the population. There are other relationships among items that are not explained by the common factor (e.g., correlated errors or minor factors). If the size of the unexplained relationships is not high (which is usually quantified by model evaluation methods in structural equation modeling; e.g., see West, Taylor, & Wu, 2012), a one-factor CFA model is assumed to be a

Table 13
The Coverage Rates of Confidence Intervals for Categorical Omega Using Four Bootstrap Methods Classified by the Number of Items and Population Reliability for Study 3

		RELIA		
Methods	Number of items	.7	.8	.9
BSE	4	.931	.929	.930
	8	.923	.906	.870
	12	.937	.902	.830
BSE-L	4	.939	.936	.945
	8	.941	.931	.906
	12	.952	.930	.879
PER	4	.931	.925	.928
	8	.907	.861	.741
	12	.919	.838	.603
BC*a*	4	.927	.929	.937
	8	.960	.956	.940
	12	.970	.961	.908

Note. L = logistic-transformed Wald confidence interval; BSE = bootstrap standard error; PER = percentile bootstrap; BC*a* = bias-corrected-and-accelerated bootstrap; RELIA = population categorical omega for perfectly fitting model.

parsimonious explanation of the population. However, coefficient omega does not account for the existence of correlated errors that are not included in the model because the formula for coefficient omega relies on parameter estimates derived from the CFA model. Coefficient omega could overestimate or underestimate the population reliability if errors are correlated. Instead of basing all estimates on the assumed correctly fitting CFA model, hierarchical omega calculates the total observed variance from the observed scores directly (i.e., the variance of *Y*). Thus, we recommend that researchers use hierarchical omega, a coefficient that builds on others in the literature that has more relaxed assumptions.[17]

For categorical items, we encourage using categorical omega because it does not require a perfectly fitting CFA model and it accounts for the appropriate relationship between the factor and categorical items. The CFA model for categorical items uses a probit link function to model the theorized underlying continuous scale into ordered categorical items. Coefficient omega and hierarchical omega both assume a linear relationship between items and factors, which is not valid when items are categorical.

Additionally, and importantly, we encourage researchers to report not only the point estimate of the reliability coefficient, but also the confidence interval for the population reliability. The confidence interval limits, as well as width, should be considered when discussing the reliability of a composite measure. Wide confidence intervals illustrate the uncertainty with which a population value has been estimated (e.g., Kelley & Maxwell, 2003; Kelley & Preacher, 2012; Maxwell, Kelley, & Rausch, 2008). Even if a composite measure has a high estimate for reliability, its confidence interval may be so wide that the population reliability could actually be very low. Without this knowledge, which is directly communicated with a confidence interval, users of the estimate are left unaware of the plausible range of parameter values (i.e., the values contained within the confidence interval). Thus, confidence intervals for reliability coefficients, we believe, are always necessary.[18]

The issue that remains once a researcher decides that he or she wants to report a confidence interval for the population reliability value is, "which confidence interval procedure should be used—there are many?" This article seeks to shed light on this question by comparing more than 10 previously proposed methods of confidence interval formation for reliability coefficients. In particular, we evaluated 12 confidence interval formation procedures (see Table 1) for coefficient alpha, coefficient omega, hierarchical omega, and categorical omega in a variety of realistic situations, such as both CFA models with and without model error, and both CFA models with continuous and categorical items. Our study is the most comprehensive study that examined topics in which we are aware, with regards to the number of confidence interval methods for four types of reliability measures across a wide variety of realistic situations. Reliability is one of the central tenants of psychometrics and the use of psychometric measures that are composite scores should always come with an estimate of reliability and the corresponding confidence interval. Without a clear rationale for knowing which of the competing methods should, or should not, be used to form a confidence interval for population reliability coefficients, researchers are at a loss for which of the many methods to use.

Throughout this article, the treatment of reliability is focused on homogeneous measurement instruments. If a measurement instrument measures multiple constructs (i.e., it is heterogeneous), the present article still is useful because the set of items measuring each construct may themselves be homogeneous (i.e., for purposes of using the items as their own homogenous scale). For example, a scale may measure three constructs and thus, by definition, not be homogeneous. However, a different composite score could be used for each of the three subscales and thus our discussion of homogeneous composites would apply to each of the three scales individually.

The simulation studies were designed to consider the effectiveness of different methods of confidence intervals from different types of models and to provide comparisons. In Study 1, CFA model without model error was considered. We found that confidence intervals for coefficient alpha did not perform well in most conditions compared to confidence intervals for coefficient omega. In particular, with the exception of the Fisher and ADF approaches, the other methods worked reasonably well and similarly to one another for normal items. For nonnormal items, only normal-theory method with robust maximum likelihood and all bootstrap methods performed well.

The results of comparing the effectiveness of different approaches for coefficient alpha and coefficient omega partially support the previous simulation studies on the confidence interval for coefficient alpha. Although previous studies (e.g., Cui & Li, 2012; Duhachek & Iacobucci, 2004; Maydeu-Olivares, Coffman, & Hartmann, 2007; Padilla et al., 2012; Romano et al., 2010) recommended some types of confidence interval for coefficient alpha, it was for estimating the population coefficient alpha, which is not the target parameter. That is, the previous studies evaluated if confidence intervals for coefficient alpha bracketed population coefficient alpha—not if the confidence intervals bracketed population reliability. Population coefficient alpha and population reliability are not generally the same. Therefore, we do not recommend using coefficient alpha and its confidence interval for estimating population reliability. Regarding coefficient omega, normal-theory method with robust maximum likelihood and the bootstrap methods provided the best performance regardless of the item distributions. Using the logistic transformation approach improved the coverage rates for both normal-theory method with robust maximum likelihood and bootstrap standard error.

The normal-theory method did not perform well for nonnormal data because maximum likelihood estimation assumes normal item distributions. Using robust maximum likelihood much improved the coverage rates of population reliability for nonnormal distri-

[17] Coefficient omega requires a just-identified latent variable. Hierarchical omega requires overidentified latent variable to have a potentially different value from coefficient omega. A just-identified latent variable cannot distinguish between the target construct and minor constructs, which is an issue that arises with categorical omega. To separate minor factors, an overidentified latent variable is needed.

[18] Terry and Kelley (2012) develop sample size planning methods in order to obtain narrow confidence intervals when using coefficients alpha or omega. These sample size planning methods can be thought of as approximations to the bootstrap confidence intervals for hierarchical omega and are available in MBESS. The methods of Terry and Kelley (2012) are approximations in this context because Terry and Kelley (2012) did not consider hierarchical omega or categorical omega and they only considered one approach each for forming confidence intervals based on coefficient alpha or coefficient omega. However, with an a priori Monte Carlo simulation study to evaluate sampling characteristics for a particular condition, the simulation-based approach they consider (specifically in section 5.1.3 of Terry & Kelley, 2012) could be extended to the situation of interest for any coefficient for any confidence interval method.

bution. The Fisher's method tended to provide overcoverage (because of widths that were larger than necessary). The ADF method's coverage was far lower than.95 in this simulation study for coefficient omega. The reason is that the estimated coefficient omega by the ADF method has larger bias than the coefficient omega estimated from the maximum likelihood (see the online supplement), which is used to find the confidence interval for other approaches. This result is consistent to Olsson, Foss, Troye, and Howell's (2000) suggestion that the ADF estimator will provide parameter estimates close to the maximum likelihood estimator only when the sample size is extremely large (e.g., ≥ 2,000). The sample size used in this simulation study was apparently not large enough to use the ADF approach for coefficient omega. The likelihood-based approach is not recommended for the formation of the confidence interval for coefficient omega for nonnormal items. Similar to Padilla and Divers (2013a) and Padilla and Divers (2013b), all bootstrap approaches performed well. Because data in applied research rarely follow normal distribution, these methods are preferred.

Although previous simulation studies supported coefficient omega, they created data from perfectly fitting models. In practice, data do not usually come from a population in which the one-factor CFA model fits perfectly. Although researchers do not know whether the misfit is attributed to sampling error or population model error, it is likely that population model error exists. All models are designed to be a simplified explanation of reality. Assuming that a model fits perfectly, and therefore using coefficient omega, is not ideal. Thus, in Study 2, we generated data with population model error. As shown in Study 2, confidence intervals for coefficient omega did not perform as well as hierarchical omega, especially when sample size is low and population reliability is high. Although the percentile bootstrap had the best coverage rates, it had only a slight difference from other bootstrap methods. A discussion on the particular method for computing a confidence bootstrap confidence interval is less pressing than actually computing a bootstrap confidence interval for the population reliability coefficient, so long as the bootstrap methods for hierarchical omega are used.

Previous simulation studies defined population reliability of categorical items by using coefficient omega. For example, Maydeu-Olivares et al. (2007) calculated observed item covariances and used a one-factor CFA model for continuous items to fit the observed item covariances and calculate coefficient omega. This coefficient omega was used as the population reliability. Coefficient omega does not represent population reliability in this situation, however, because the relationship between the factor and items is not linear. Recently, Green and Yang (2009a) proposed a method of calculating population reliability for categorical items. Study 3 evaluated the performances of hierarchical omega and categorical omega for models with categorical items accounting for the nonlinear relationship between items and factor. In this simulation study, we also simulated data that does not fit the one-factor model perfectly, as in Study 2, mimicking what we consider to be realistic scenarios in practice. We found that confidence intervals for hierarchical omega performed well only when threshold patterns were similar across all items. When threshold patterns were different across items, hierarchical omega had a different value from categorical omega, so the coverage rates were not acceptable. However, BC*a* confidence intervals for categorical omega had good coverage rates in all threshold patterns, thus the BC*a* method is recommended for categorical items.

These simulation studies showed the effectiveness of the different approaches to confidence interval formation for coefficient alpha, coefficient omega, hierarchical omega, and categorical omega in real-world situations. Although this simulation cannot be generalized to all situations that are of potential interest, we hope that this article will help researchers understand implications of using different reliability coefficients and different confidence interval formation methods for the population reliability. The choice of confidence interval approach is very influential for appropriate confidence interval coverage (i.e., of obtaining an empirical coverage rate similar to the nominal coverage rate).

In conclusion, we have several recommendations. First, avoid using coefficient alpha and thereby computing confidence intervals for coefficient alpha. We understand that in many settings using coefficient alpha is expected, if not literally required, when using composite scores. From a practical standpoint, if coefficient alpha is being required, it can be included along with a more appropriate reliability coefficient. We suggest avoiding coefficient alpha as an estimate of the population reliability because it only estimates the population reliability under the rather unrealistic assumption that true-score equivalence holds; otherwise it provides a systematic underestimate of the population reliability (in situations in which classical test assumptions hold). Coefficient omega has its own potentially unrealistic limitation, namely that it assumes a perfectly fitting single factor model. In situations in which minor factors (or correlated errors) exist, coefficient omega does not estimate population reliability. Therefore, our second recommendation is, when items are continuous, estimate population reliability with hierarchical omega. Hierarchical omega provides a generalization of coefficient omega by using the variance of the composite as the denominator, which maps on perfectly to the idea of reliability being defined as the "true variance" to the "total variance." Unless the model fits the one factor model perfectly, coefficient omega provides an incorrect value for the total variance. Our third recommendation is for categorical items, where we recommend categorical omega. Fourth, we recommend any of the bootstrap confidence interval procedures discussed when forming confidence intervals for population reliability using hierarchical omega as an estimate. If coefficient omega is used, we also recommend bootstrap confidence intervals. Fifth, we recommend using BC*a* confidence intervals for categorical omega. Additionally, we have provided software to implement all of the estimation and confidence interval procedures with the `ci.reliability()` function provided in the MBESS Kelley (2016, 2007b, 2007a) R (R Development Core Team, 2015) package.

References

American Psychological Association. (2010). *Publication manual of the American psychological association* (6th ed.). Washington, DC: American Psychological Association.

Association for Psychological Science. (2014). *2014 submission guidelines*. Retrieved from http://www.psychologicalscience.org/index.php/publications/journals/psychological/_science/ps-submissions

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

Bentler, P. M., & Satorra, A. (2010). Testing model nesting and equivalence. *Psychological Methods, 15,* 111–123.

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., . . . Fox, J. (2011). OpenMx: An open source extended structural equation modeling framework. *Psychometrika, 76,* 306–317.

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics, 27,* 335–340.

Bonett, D. G. (2010). Varying coefficient meta-analytic methods for alpha reliability. *Psychological Methods, 15,* 368–385.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31,* 144–152.

Browne, M. W. (1982). Covariance structures. In D. M. Hawkins (Ed.), *Topics in applied multivariate analysis* (pp. 72–141). Cambridge, UK: Cambridge University Press.

Browne, M. W. (1984). Asymptotic distribution free methods in the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 24,* 445–455.

Cheung, M. W.-L. (2009). Constructing approximate confidence intervals for parameters with structural constructing approximate confidence intervals for parameters with structural equation models. *Structural Equation Modeling, 16,* 267–294.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth.

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

Cudeck, R., & Browne, M. W. (1992). Constructing a covariance matrix that yields a specified minimizer and a specified minimum discrepancy function value. *Psychometrika, 57,* 357–369.

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

Duhachek, A., & Iacobucci, D. (2004). Alpha's standard error (ASE): An accurate and precise confidence interval estimate. *Journal of Applied Psychology, 89,* 792–808.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY: Chapman & Hall/CRC.

Enders, C. K. (2001). The impact of nonnormality on full information maximum likelihood estimation for structural equation models with missing data. *Psychological Methods, 6,* 352–370.

Feldt, L. S. (1965). The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika,* 357–370.

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement, 11,* 93–103.

Fisher, R. A. (1950). *Statistical methods for research workers*. Edinburgh, UK: Oliver & Boyd.

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.

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods, 19,* 72–91.

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

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

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

Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley.

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

Hakstian, A. R., & Whalen, T. E. (1976). A *k*-sample significance test for independent alpha coefficients. *Psychometrika, 41,* 219–231.

Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika, 6,* 153–160.

Iacobucci, D., & Duhachek, A. (2003). Advancing alpha: Measuring reliability with confidence. *Journal of Consumer Psychology, 13,* 478–487.

Jöoreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36,* 109–133.

Kelley, K. (2007a). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software, 20,* 1–24.

Kelley, K. (2007b). Methods for the Behavioral, Educational, and Educational Sciences: An R package. *Behavior Research Methods, 39,* 979–984.

Kelley, K. (2016). MBESS (Version 4.0.0) [computer software and manual], retrieved from http://cran.r-project.org

Kelley, K., & Cheng, Y. (2012). Estimation of and confidence interval formation for reliability coefficients of homogeneous measurement instruments. *Methodology, 8,* 39–50.

Kelley, K., & Maxwell, S. E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods, 8,* 305–321.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods, 17,* 137–152.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: Guilford Press.

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

Koning, A. J., & Franses, P. H. (2003). *Confidence intervals for Cronbach's coefficient alpha values* (Tech. Rep. No. ERIM Report Series Ref. No. ERS-2003-041-MKT). Rotterdam, the Netherlands: Erasmus Research Institute of Management.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2,* 151–160.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods, 13,* 203–229.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology, 51,* 201–226.

Marcoulides, G. A., & Saunders, C. (2006). PLS: A silver bullet? *Management Information Systems Quarterly, 30,* iii–ix.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology, 59,* 537–563.

Maydeu-Olivares, A., Coffman, D. L., & Hartmann, W. M. (2007). Asymptotically distribution-free (ADF) interval estimation of coefficient alpha. *Psychological Methods, 12,* 157–176.

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

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

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39,* 479–515.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49,* 115–132.

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

Oehlert, G. W. (1992). A note on the delta method. *The American Statistician, 46,* 27–29.

Olsson, U. H., Foss, T., Troye, S. V., & Howell, R. D. (2000). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling, 7,* 557–595.

Padilla, M. A., & Divers, J. (2013a). Bootstrap interval estimation of reliability via coefficient omega. *Journal of Modern Applied Statistical Methods, 12,* 79–89.

Padilla, M. A., & Divers, J. (2013b). Coefficient omega bootstrap confidence intervals nonnormal distributions. *Educational and Psychological Measurement,* 956–792.

Padilla, M. A., Divers, J., & Newton, M. (2012). Coefficient alpha bootstrap confidence interval under nonnormality. *Applied Psychological Measurement, 36,* 331–348.

Rasch, D., Kubinger, K. D., & Moder, K. (2011). The two-sample *t* test: Pre-testing its assumptions does not pay off. *Statistical, 52,* 219–231.

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

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

Raykov, T. (2002b). Automated procedure for obtaining standard error and confidence interval for scale reliability. *Understanding Statistics, 1,* 75–84.

Raykov, T. (2012). Scale construction and development using structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 472–492). New York, NY: Guilford Press.

Raykov, T., Dimitrov, D. M., & Asparouhov, T. (2010). Evaluation of scale reliability with binary measures using latent variable modeling. *Structural Equation Modeling, 17,* 265–279.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York, NY: Routledge.

Raykov, T., & Marcoulides, G. A. (2013). Meta-analysis of scale reliability using latent variable modeling. *Structural Equation Modeling, 20,* 338–353.

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

R Development Core Team. (2015). *R: A language and environment for statistical computing* [Computer software manual].

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

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

Romano, J. L., Kromrey, J. D., & Hibbard, S. T. (2010). A Monte Carlo study of eight confidence interval methods for coefficient alpha. *Educational and Psychological Measurement, 70,* 376–393.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48,* 1–36.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66,* 507–514.

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

Siotani, M., Hayakawa, T., & Fujikoshi, Y. (1985). *Modem multivariate statistical analysis: A graduate course and handbook*. Columbus, OH: American Sciences Press.

Task Force on Reporting of Research Methods in AERA Publications. (2006). *Standards for reporting on empirical social science research in AERA publications, American educational*. Washington, DC: American Educational Research Association.

Terry, L. J., & Kelley, K. (2012). Sample size planning for composite reliability coefficients: Accuracy in parameter estimation via narrow confidence intervals. *British Journal of Mathematical and Statistical Psychology, 65,* 371–401.

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika, 48,* 465–471.

van Zyl, J. M., Neudecker, H., & Nel, D. G. (2000). On the distribution of the maximum likelihood estimator of Cronbach's alpha. *Psychometrika, 65,* 271–280.

Venables, W. N., & Ripley, B. D. (2015). *MASS* [Computer software and manual]. Retrieved from http://www.cran.r-project.org/

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209–231). New York, NY: Guilford Press.

Woodhouse, B., & Jackson, P. H. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: II: A search procedure to locate the greatest lower bound. *Psychometrika, 42,* 579–591.

Yuan, K.-H., Guarnaccia, C. A., & Hayslip, B., Jr. (2003). A study of the distribution of sample coefficient alpha with the Hopkins Symptom Checklist: Bootstrap versus asymptotics. *Educational and Psychological Measurement, 63,* 5–23.

Zimmerman, D. W. (1975). Probability spaces, Hilbert spaces, and the axioms of test theory. *Psychometrika, 40,* 395–412.

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

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

(*Appendix follows*)

Appendix

Coefficient Omega and the Variance of a Composite

The goal of this appendix is to show how to derive the formula of coefficient omega. First, recall that the population coefficient omega was given as

$$\omega = \frac{\left(\sum_{j=1}^{J} \lambda_j\right)^2}{\left(\sum_{j=1}^{J} \lambda_j\right)^2 + \sum_{j=1}^{J} \psi_j^2}. \text{ (Equation 11, repeated)}$$

Second, recall that the observed score on the *j*th component for the *i*th individual from Equation 1 is

$$X_{ij} = T_{ij} + \epsilon_{ij}. \text{ (Equation 1, repeated)}$$

Third, recall that the true score on an item for an individual (from Equation 1) can be represented in the factor analytic perspective as

$$T_{ij} = \mu_j + \lambda_j \eta_i. \text{ (Equation 10, repeated)}$$

We denote the $\lambda_j \eta_i$ part of Equation 10 (i.e., the true part and shown above) as τ_{ij}:

$$\tau_{ij} = \lambda_j \eta_i. \quad (40)$$

Thus, Equation 10 can be rewritten as

$$T_{ij} = \mu_j + \tau_{ij}. \quad (41)$$

Over items for the *i*th individual the second component of the right-hand-side of Equation 41 generalizes to

$$\tau_i = \sum_{j=1}^{J} \tau_{ij} = \left(\sum_{j=1}^{J} \lambda_j\right) \eta_i, \quad (42)$$

and represents the true part of the *i*th individual's composite, with the variance of τ_i across individuals denoted σ_τ^2 (i.e., true variance). The error of individual *i*'s composite score is

$$\epsilon_i = \sum_{j=1}^{J} \epsilon_{ij}, \quad (43)$$

with the variance of ϵ_i across individuals denoted σ_ϵ^2 (i.e., error variance). That is, Equation 2 (the composite for the *i*th individual) can be rewritten in a factor analytic framework as

$$Y_i = \sum_{j=1}^{J} \mu_j + \tau_i + \epsilon_i. \quad (44)$$

Further, consider the variance of *y* written in the factor analytic framework:

$$\text{Var}(Y) = \text{Var}\left(\sum_{j=1}^{J} \mu_j + \tau_i + \epsilon_i\right) \quad (45)$$

$$= \text{Var}\left(\sum_{j=1}^{J} \mu_j\right) + \text{Var}(\tau_i) + \text{Var}(\epsilon_i), \quad (46)$$

which reduces to

$$\text{Var}(Y) = \text{Var}(\tau_i) + \text{Var}(\epsilon_i) \quad (47)$$

because $\sum_{j=1}^{J} \mu_j$ is a constant. Equation 47 can be rewritten (by replacing τ_i with $\eta_i \sum_{j=1}^{J} \lambda_j$, from Equation 40) as

$$\text{Var}(Y) = \text{Var}\left(\eta_i \sum_{j=1}^{J} \lambda_j\right) + \text{Var}(\epsilon_i). \quad (48)$$

Due to the errors being uncorrelated and letting ψ_j^2 denotes the error variance of the *j*th item, $\text{Var}(\epsilon_i)$ can be written as the sum of the variances:

$$\text{Var}(Y) = \text{Var}\left(\eta_i \sum_{j=1}^{J} \lambda_j\right) + \sum_{j=1}^{J} \psi_j^2. \quad (49)$$

Because $\sum_{j=1}^{J} \lambda_j$ is a constant, Equation 49 reduces to:

$$\text{Var}(Y) = \left(\sum_{j=1}^{J} \lambda_j\right)^2 \text{Var}(\eta_i) + \sum_{j=1}^{J} \psi_j^2. \quad (50)$$

Recalling that we fixed $\text{Var}(\eta_i)$ to 1 for model identification purposes (i.e., it is a constant), the variance of the composite is thus

$$\text{Var}(Y) = \left(\sum_{j=1}^{J} \lambda_j\right)^2 + \sum_{j=1}^{J} (\psi_j^2). \quad (51)$$

Notice that the first component on the right-hand-side of Equation 51 is the variance of *Y* due to the model (i.e., the true part), whereas the second component on the right-hand-side of Equation 51 is the variance of *Y* due to the errors. These two variances are how the terms in Equation 6 can be operationalized in a factor analytic conceptualization of reliability (e.g., McDonald, 1999).

Received December 21, 2012
Revision received September 20, 2015
Accepted November 6, 2015 ■