Original Article

Estimation of and Confidence Interval Formation for Reliability Coefficients of Homogeneous Measurement Instruments

Ken Kelley[1] and Ying Cheng[2]

[1]Department of Management, University of Notre Dame, IN, USA, [2]Department of Psychology, University of Notre Dame, IN, USA

Abstract. The reliability of a composite score is a fundamental and important topic in the social and behavioral sciences. The most commonly used reliability estimate of a composite score is coefficient α. However, under regularity conditions, the population value of coefficient α is only a lower bound on the population reliability, unless the items are essentially τ-equivalent, an assumption that is likely violated in most applications. A generalization of coefficient α, termed ω, is discussed and generally recommended. Furthermore, a point estimate itself almost certainly differs from the population value. Therefore, it is important to provide confidence interval limits so as not to overinterpret the point estimate. Analytic and bootstrap methods are described in detail for confidence interval construction for ω. We go on to recommend the bias-corrected bootstrap approach for ω and provide open source and freely available R functions via the MBESS package to implement the methods discussed.

Keywords: reliability, coefficient α, Cronbach's α, coefficient ω, confidence intervals, bootstrap technique

Researchers have a vested interest in the quality of the measurement instruments they employ. One important, but sometimes overlooked, aspect of a research study is the reliability of the scores upon which inferences will be based and the proper way the reliability of scores should be reported and interpreted. Reliability has been a central issue in the measurement of constructs for more than a century, dating to at least Spearman when he attempted to overcome attenuated correlations due to observational errors that "inevitably arise from errors in measuring" (Spearman, 1904, p. 253; see Jones & Thissen, 2007, for a review). Although measurement error is ubiquitous in research, better measurement leads to better inferences and ultimately better conclusions. Without high-quality measurements, the conclusions and inferences based on those measurements can be called into question and, in some circumstances, shown to be misleading or even incorrect. If the integrity of conclusions and inferences cannot be trusted because of unreliable scores in a particular study, the study will likely contribute little to the cumulative knowledge of a discipline.

In a study utilizing composite scores from a certain measurement instrument collected on a sample, whether the instrument is a test, procedure, questionnaire, survey, rating scale, etc., it is a general recommendation that the reliability of these scores be reported along with the information about the population from which the sample was selected. Although researchers using established measurement instruments with known reliability properties are encouraged to report the reliability as provided by the developer, the developer reported reliability may or may not be the same as for the particular population from which the sample was selected (e.g., a clinical sample, a different age group, a sample from another country, a test translated into another language, etc.). It is important to realize that a measurement instrument is itself not inherently reliable or unreliable and that reliability is not an inherent property of a particular measurement instrument – rather, reliability is a property of the scores for a particular population on a particular measurement instrument (e.g., Thompson, 2003; Wilkinson & the American Psychological Association Task Force on Statistical Inference, 1999).

Arguably more important than the estimated reliability coefficient itself, however, is the confidence interval for the population value of the reliability coefficient. The confidence interval is valuable because the point estimate itself almost certainly does not equal the population parameter exactly, however, providing the confidence limits identifies a range of plausible parameter values that will contain the

Methodology 2012; Vol. 8(2):39–50
DOI: 10.1027/1614-2241/a000036

population value of the reliability with the specified (e.g., 90%, 95%, or 99%) degree of confidence.[1] It is this population value that is of interest, not the point estimate itself. As Wilkinson and the American Psychological Association (APA) Task Force on Statistical Inference recommended, "*interval estimates should be given for any effect size involving principal outcomes*" (1999, p. 599). In the context of measurement instruments, the effect size of interest is the reliability coefficient of the scores. In line with the recommendation from Wilkinson and the APA Task Force, some journals request or require confidence intervals of reliability coefficients whenever these coefficients are reported in their instructions to authors (e.g., Fan & Thompson, 2001).

This article assumes a homogeneous measurement instrument, which is an instrument that measures only a single construct, where a one-factor model will hold. The standard procedure for estimating the reliability of a homogeneous measurement instrument in the social and behavioral sciences is coefficient α. However, given the single construct measured in a homogeneous measurement instrument, the population value of coefficient α is only a lower bound on the population value of reliability, unless the stringent assumption of essential τ-equivalence holds, which is an assumption that is likely untenable in most applied settings (Novick & Lewis, 1967).[2] The assumption of essential τ-equivalence is such that item score means differ only by additive constants, but not in scale (Lord & Novick, 1968, p.50). Under the one-factor model, this means the items are equally sensitive (i.e., same value of the factor loading) at measuring the underlying construct. Essential τ-equivalence has also been referred to in the literature as a true-score equivalence model (e.g., McDonald, 1999, p. 85). An alternative estimate of population reliability, termed omega (ω), allows for some items to be more or less sensitive than others at measuring the underlying construct. A formal discussion clarifying the similarities and differences of coefficient α and ω is forthcoming.

One thing that is clear from the applied literature is that coefficient α is currently the way in which researchers tend to operationalize the reliability of a measurement instrument. However, the meaning and usage of coefficient α is such a pressing matter that *Psychometrika* recently published an article calling into question the usefulness of coefficient α (Sijtsma, 2009a) with commentary articles (Bentler, 2009; Green & Yang, 2009a, 2009b; Revelle & Zinbarg, 2009) and a rejoinder (Sijtsma, 2009b). We second many concerns raised in those recent articles. We address the concerns by (a) advocating the use of an appropriate reliability coefficient, (b) discussing the importance of a confidence interval for the population value of reliability, (c) reviewing methods of confidence interval formation for population reliability coefficients, and (d) offering an easy-to-use R package so that applied researchers can use the methods we discuss.

In particular, we argue that coefficient α is oftentimes an inadequate estimate of reliability and that there are other better estimates available, including the one that will be discussed at length in this paper, ω. Although other estimates of reliability are available (e.g., maximal reliability, φ [Li, 1997; Yuan & Bentler, 2002], Revelle's β [1979], greatest lower bound [Ten Berge, 2004], and reliability for a hierarchical factor, ω_h [McDonald, 1970; Zinbarg, Revelle, Yovel, & Li, 2005]), our discussion focuses on ω because it is the best reliability estimate for composite scores of homogeneous tests.[3] Therefore, we provide a tutorial type treatment of ω and emphasize that a point estimate is not enough – a confidence interval for the population reliability coefficient is needed also to effectively communicate information about the reliability of a particular instrument in a particular population. Correspondingly, we discuss the construction of confidence intervals for population ω using both analytic and bootstrap methods. As Revelle and Zinbarg (2009) have noted, quality software must exist and be available to those who would benefit from its use. Along those lines, we provide a freely available and easy-to-use software package (Kelley, 2007a, 2007b; Kelley & Lai, 2010) that is part of a more general and freely available software program (R Development Core Team, 2010). We demonstrate the methods discussed with the software provided in order to illustrate how the software can be used. We believe that the expository nature of this article, coupled with the software that we have provided, will promote the use of the methods we discuss and will be helpful for advancing social and behavioral science.

Estimating the Reliability of an Unweighted Composite

The classical test theory (CTT) decomposition of the observed value for a particular item is

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

where X_{ij} is the observed value for the ith individual ($i = 1, ..., N$) on the jth item ($j = 1, ..., J$), τ_{ij} is the true-score for the ith individual on the jth item, and ϵ_{ij} is the error for the ith individual on the jth item (e.g., Guilford, 1954; Gulliksen, 1950; Lord & Novick, 1968; McDonald, 1999; Zimmerman, 1975).[4] The theorem defining CTT

[1] The confidence interval is based on random data and is thus itself a random value. Provided assumptions are met, if an infinite number of confidence intervals were computed, $(1 - \alpha')100\%$ of the confidence intervals would bracket the value of the population parameter, where $(1 - \alpha')$ is the desired confidence level. See Hahn and Meeker (1991) for more details about the technical meaning of confidence intervals.

[2] When the one-factor model does not hold because errors are correlated and are treated purely as measurement errors, coefficient α may overestimate the value of the population reliability (Green & Hershberger, 2000; Komaroff, 1997; Zimmerman, Zumbo, & Lalonde, 1993). Throughout the article, we assume that the errors are uncorrelated.

[3] When a general factor can be identified for a scale that measures multiple constructs, a hierarchical version of ω (ω_h) will be more appropriate (Zinbarg et al., 2005). When only one construct is being measured, as the focus of this paper, ω is the most appropriate.

[4] It should be noted that it is assumed that the variance of the X_{ij} values and the variance of ϵ_{ij} values are finite, which further implies that the means are finite, since if the $r + 1$ moment is finite so too must be the rth moment (e.g., Lord & Novick, 1968, p. 36).

states that the errors of measurement (i.e., the $\epsilon_{\cdot j}$s) are mutually uncorrelated (i.e., $\rho(\epsilon_{\cdot j}, \epsilon_{\cdot j'}) = 0$ for all $j \neq j'$), are uncorrelated with all true-scores (i.e., $\rho(\tau_j, \epsilon_{\cdot j'}) = 0$ for all $j = j'$ and $j \neq j'$), and have a mean of zero (i.e., $E[\epsilon_{\cdot j}] = 0$), 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). With the appropriate generalization of the models, the stated assumptions can be relaxed and a more general framework can be used.

Although Equation 1 is the CTT representation of an observed item, an individual's score for many measurement instruments is the sum of the values from the J items. Forming a score from the sum of items generally makes sense only when the items form a homogeneous measurement instrument, which is an instrument that measures only a single construct. For homogeneous measurement instruments, the score for the measurement instrument is termed an unweighted (unit-weighted) composite,

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

Because Y_i is itself an observed score, it can be conceptualized in a form analogous to Equation 1,

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

where $\tau_i = \sum_{j=1}^{J} \tau_{ij}$ and $\epsilon_i = \sum_{j=1}^{J} \epsilon_{ij}$. Note that there are no j subscripts for Y or τ in Equation 3, as these values represent the observed (composite) score and the true (composite) score, respectively, for the ith individual.

The psychometric definition of reliability for an unweighted (unit-weighted) composite is

$$\rho(Y) = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\epsilon^2}, \tag{4}$$

which can be rewritten as

$$\rho(Y) = \frac{\sigma_\tau^2}{\sigma_Y^2}, \tag{5}$$

where σ_τ^2 is the population variance of the true-scores for the composite (i.e., τ_i values), σ_ϵ^2 is the population variance of the error of the scores for the composite (i.e., ϵ_i values), and σ_Y^2 is the population variance of the observed scores for the composite (i.e., Y_i values). Notice that σ_Y^2 is the sum of σ_τ^2 and σ_ϵ^2, which holds because of the assumption that τ and ϵ are uncorrelated. In other words, the psychometric definition of reliability is the ratio of the variance of the true-scores to the sum of the variance of the true-scores and the variance of the errors, or equivalently, the ratio of the variance of the true-scores to the variance of the observed scores. Because unweighted (unit-weighted) composite scores play an important role in the literature of the social and behavioral sciences, understanding issues of reliability of composite scores is important for researchers who use such scales directly or indirectly.

Coefficient α

In 1951 Cronbach authored a conceptually appealing treatment of reliability in a manner that was accessible to many researchers.[5] The definition of coefficient α is given as

$$\alpha \equiv \frac{J}{J-1}\left(1 - \frac{\sum_{j=1}^{J} \sigma_j^2}{\sigma_Y^2}\right), \tag{6}$$

where σ_j^2 denotes the population variance of the jth item. It can be shown that coefficient α is a "lower bound" on the true reliability of a set of scores, where it underestimates reliability under the usual CTT assumptions unless items are essentially τ-equivalent (see, e.g., Lord & Novick, 1968, pp. 87–90; McDonald, 1999, pp. 92–93; Novick & Lewis, 1967). Recall that essentially τ-equivalent items are items that have means that differ only by additive constants.[6]

McDonald (1999) shows that a general factor analytic model can be used as a way to largely unify seemingly diverse models in CTT. Figure 1 is a path diagram, using reticular action model notation (McArdle & McDonald, 1984) of a true-score equivalent model from a factor analytic perspective, where η is the common factor that the J items are thought to measure. Without loss of generality, we assume the variance of the common factor to be 1 (i.e., $\sigma_\eta^2 = 1$) . Figure 1 shows that the true-score for the ith individual on any item is the product of the individual's factor score and the factor loading as

$$\tau_{ij} = \eta_i \lambda. \tag{7}$$

Notice from Equation 7 that a j subscript is not necessary for λ because they are constant across all items for all individuals. Figure 1 makes it explicit that the underlying factor (i.e., η) is measured by each of the J items (i.e., the X_j values) with equal sensitivity, but with potentially unique error variances (i.e., the ψ_j^2 values) of the item errors (i.e., the ϵ_j values).

Recall that in the special case of essential τ-equivalence, coefficient α is exactly equal to the composite reliability [i.e., $\alpha = \rho(Y)$]. However, because true-score equivalence is not likely to hold in many situations in applied research, coefficient α is not recommended here as an estimate of composite reliability. Historically, coefficient α has been the most widely recommended and used measure of reliability of the social and behavioral sciences. From a theoretical perspective, however, the assumption that all items measure

[5] Note that Equation 6 was published as L_3 by Guttman (1945, p. 259) 6 years before Cronbach published the same formula that he denoted α. Guttman's L_3 is a generalization of Kuder and Richardson's Equation 20 (KR-20) (1937, p. 158) when items are continuous. Rather than referring to Equation 6 as "Cronbach's α," which is often done in the literature to Cronbach's embarrassment (Cronbach & Shavelson, 2004, p. 397), it is referred to as coefficient α, which was Cronbach's original intent and preference (Cronbach & Shavelson, 2004).

[6] Following the factor analytic model, essentially τ-equivalent and τ-equivilant also means that the item covariances are all equal.

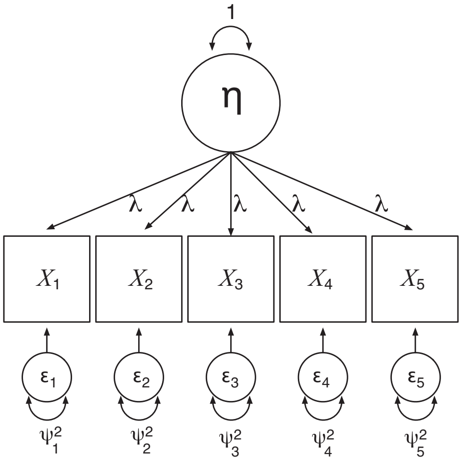

Figure 1. Path diagram for a homogeneous measurement instrument, where the underlying attribute (η) has been measured by five items (X_1–X_5), and where essential τ-equivalence (i.e., true-score equivalence) holds.

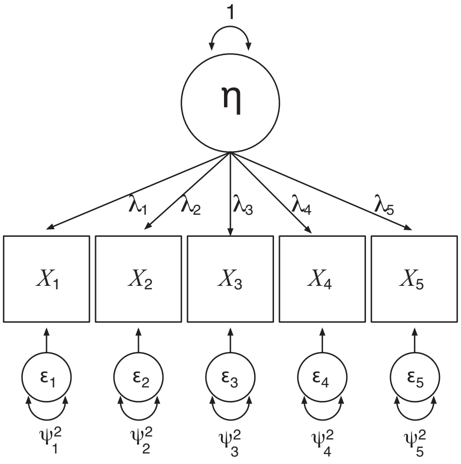

Figure 2. Path diagram for a homogeneous measurement instrument, where the underlying attribute (ξ) has been measured by five items (X_1–X_5), and where essential τ-equivilance (i.e., true-score equivalence) does not hold.

the factor equally sensitive is almost certainly false in many applications. Thus, basing the estimated reliability of a set of scores on a reliability coefficient that is thought to be inappropriate a priori in the vast majority of situations is problematic. The next section describes a generalization of coefficient α to the situation where the items need not measure the factor with the same degree of sensitivity, which implies that the assumption of all factor loadings being equal is relaxed.

Coefficient ω

Coefficient α makes the assumption of essential τ-equivalence, in other words, equal population factor loadings, which leads to a theoretically implausible model in many practical situations. Here we introduce a model that allows factor loadings to differ (see Figure 2). This model can easily be used for estimating the reliability of a set of scores for some measurement instrument. In Figure 2, λ_j is the path coefficient linking the underlying attribute (η) to the observed score (X_j), and ϵ_j is the error of the particular observed score with mean zero and variance ψ_j^2. (Note that $\tau_{ij} = \eta_i \lambda_j$.)

Notice that unlike Figure 1, each path coefficient (i.e., λ value) in Figure 2 has a corresponding subscript. Figure 2 represents a flexible model that is known as congeneric. A congeneric factor model is one where a single factor has factor loadings that need not be equivalent across the items and where the error variances are potentially unique.

Although the true-score equivalent and the congeneric models are different, the definition of reliability remains the same, which is that reliability is the ratio of the true-score variance to the observed variance. However, the models have fundamentally different ways of estimating the true-score variance. The estimated true variance in this case, that is, the variance that is due to the homogeneous factor model, is a function of the estimated model parameters. It can be shown that for a congeneric model, the variance of Y is

$$Var(Y) = Var\left(\sum_{j=1}^{J} \lambda_j \eta\right) + Var\left(\sum_{j=1}^{J} \epsilon_j\right). \tag{8}$$

Recalling that the error variances are uncorrelated with one another, the variance of the sum of the errors is simply the sum of the error variances, implying

$$Var(Y) = Var\left(\sum_{j=1}^{J} \lambda_j \eta\right) + \sum_{j=1}^{J} \psi_j^2. \tag{9}$$

Applying covariance algebra rules to Equation 9, with the realization that the J λ_j values are fixed in any situation, Equation 9 can be rewritten as

$$Var(Y) = \left(\sum_{j=1}^{J} \lambda_j\right)^2 Var(\eta) + \sum_{j=1}^{J} \psi_j^2. \tag{10}$$

Because *Var (η)* is a fixed quantity we set to 1, due to the necessity in factor analysis for model identification, the variance of Y is reduced to

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

Given Equation 11, the reliability coefficient for the congeneric factor model, termed coefficient ω, is defined as follows

$$\omega \equiv \frac{\left(\sum_{j=1}^{J} \lambda_j\right)^2}{\left(\sum_{j=1}^{J} \lambda_j\right)^2 + \sum_{j=1}^{J} \psi_j^2}, \tag{12}$$

which can be rewritten as

$$\omega \equiv \frac{\left(\sum_{j=1}^{J} \lambda_j\right)^2}{\sigma_Y^2}, \tag{13}$$

or

$$\omega \equiv 1 - \frac{\left(\sum_{j=1}^{J} \psi_j^2\right)}{\sigma_Y^2}. \tag{14}$$

The population value ω can be estimated by substituting the corresponding sample values for their population analogs in Equation 12 (or equivalently Equation 13 or 14). Note that the model parameters can be estimated using different methods and consequently can affect the estimation of ω. But in terms of the population value, ω is at least as large as α, given the assumptions introduced earlier.

In fact, coefficient α is a special case of coefficient ω. In particular, coefficient α and coefficient ω are equal if and only if $\lambda_1 = \lambda_2 = \cdots = \lambda_J$. In particular, by imposing such constraints on the path coefficients, coefficient α can be conceptualized as

$$\alpha = \frac{(J\lambda)^2}{(J\lambda)^2 + \sum_{j=1}^{J} \psi_j^2}. \tag{15}$$

Therefore, rather than estimating coefficient α by way of Equation 6, it can be estimated by applying a homogeneous factor model with the restriction of essential τ-equivilance (i.e., true-score equivalence), that is, setting the path coefficients (i.e., the λ_j values) to be equal. Estimating coefficient α then proceeds by using a special case of Equation 12 where the model of essential τ-equivilance (i.e., true-score equivalence) has been imposed (i.e., the λ_j values are set to a single value).

The discussion thus far has attempted to reframe existing but disparate knowledge of reliability theory that is often presented in a technical manner. As compared to the estimation of reliability coefficients, relatively little information exists on methods of confidence interval formation for coefficient ω (cf. Cheung, 2009; Raykov, 1997; Yuan & Bentler, 2002). Furthermore, such discussions tend to be rather technical with estimation procedures that are not readily known or easily implemented in standard software, thus making implementation of the methods difficult for many of the researchers who might be most interested in using them. The next two subsections illustrate three methods of confidence interval formulation for ω.

Analytic Confidence Intervals for Coefficient ω

Using the delta method, Raykov (2002) discussed an analytic confidence interval for ω that is asymptotically correct for multivariate normally distributed items, as the sample size grows toward infinity. The delta method is a way to obtain a standard error for a function of one or more parameter estimates, which can then be used for confidence interval formation in certain situations (e.g., Casella & Berger, 2002; Oehlert, 1992). The method produces asymptotically correct confidence intervals, where "asymptotically correct" refers to the confidence interval procedure actually producing $(1 - \alpha')100\%$ confidence intervals when sample size approaches infinity. This implies that for finite sample size the procedure is "approximately correct." This issue, however, is not unique to confidence intervals for reliability coefficients, but rather is generally the case for any confidence interval constructed on the basis of asymptotic theory.[7]

Using the parameters from the homogeneous congeneric factor model, let

$$\upsilon = \sum_{j=1}^{J} \lambda_j \tag{16}$$

and

$$v = \sum_{j=1}^{J} \psi_j^2. \tag{17}$$

ω (Equation 12) can then be written as

$$\omega = \frac{\upsilon^2}{\upsilon^2 + v}. \tag{18}$$

Let

$$\Delta_1 = \frac{2\upsilon v}{(\upsilon^2 + v)^2} \tag{19}$$

and

$$\Delta_2 = -\frac{\upsilon^2}{(\upsilon^2 + v)^2}, \tag{20}$$

where Δ_1 and Δ_2 are the first derivatives of Equation 18 with respect to υ and v, respectively. Defining υ and v, and then Δ_1 and Δ_2 based on υ and v, allows for easier manipulation of the parameter estimations obtained from the confirmatory factor model so that they can be used more straightforwardly in the derivation of the confidence interval procedure.

[7] Yuan and Bentler (2002) investigated the robustness of the asymptotic properties of analytic confidence intervals of several reliability coefficients, including ω, when the assumption of multivariate normality is violated. They also offered a general solution to confidence interval construction when the multivariate normality assumption is violated. When multivariate normality holds, Raykov (2002)'s method is easier to understand due to the way in which it is parameterized and is less computationally laborious. Ultimately, we suggest another approach for confidence interval construction when multivariate normality does not hold that we believe is more generally applicable, does not require asymptotically large sample sizes, and is embedded in the bootstrap method of confidence interval formation.

The approximate standard error of the estimated value of ω can be given as

$$SE(\widehat{\omega}) \approx \left[\widehat{\Delta}_1^2 Var(\widehat{\upsilon}) + \widehat{\Delta}_2^2 Var(\widehat{\nu}) + 2\widehat{\Delta}_1\widehat{\Delta}_2 Cov(\widehat{\upsilon}, \widehat{\nu})\right]^{1/2}, \tag{21}$$

where $Cov(\cdot, \cdot)$ represents the covariance of the quantities in parentheses, $\widehat{\upsilon}$, $\widehat{\nu}$, $\widehat{\Delta}_1$, and $\widehat{\Delta}_2$ are estimates of υ, ν, Δ_1, and Δ_2, respectively (Raykov, 2002). Given the standard error of $\widehat{\omega}$, an approximate $(1 - \alpha')100\%$ confidence interval can be formed as follows:

$$probability\left[\widehat{\omega} - z_{1-\alpha'/2}SE(\widehat{\omega}) \leq \omega \leq \widehat{\omega} + z_{1-\alpha'/2}SE(\widehat{\omega})\right] \approx 1 - \alpha'/2, \tag{22}$$

where $z_{1-\alpha'/2}$ is the $1-\alpha'/2$ quantile of the standard normal distribution, and $SE(\widehat{\omega})$ is equal to Equation 21 and estimated using the estimates obtained from the homogeneous factor model.

The way in which Raykov (2002) suggests the components of the standard errors from Equation 21 be estimated relies on nonlinear parameter constraints in the program LISREL (Jöreskog & Sörbom, 1996) with maximum likelihood optimization, where part of the LISREL output is then analyzed with another program or via hand calculations. Although Raykov (2002) provides example code, the relative difficulty in implementation is likely to deter many users. Our approach is one based on the same underlying maximum likelihood theory implemented in a different way. In our approach, the estimates of the components (i.e., $Var(\widehat{\upsilon})$, $Var(\widehat{\nu})$, and $Cov(\widehat{\upsilon}, \widehat{\nu})$) are easier to obtain in any fully capable structural equation model or confirmatory factor analysis program. Note that the estimates obtained from our modified implementation are equivalent to those obtained through Raykov (2002), but available without nonlinear constraints or special programing. Our modified implementation is detailed in Appendix A. Ultimately we suggest researchers use the MBESS R package where implementation is automated. Appendix B shows how the MBESS R package can easily be used. We now discuss the bootstrap methodology so that we can apply the bootstrap to the aforementioned reliability coefficients in an effort to obtain a statistically optimal confidence interval for reliability coefficients.

The General Bootstrap Technique

The bootstrap technique is a nonparametric alternative to parametric statistical techniques. The major advantage of the bootstrap technique is that it does not rely on the potentially untenable assumptions of "standard" statistical techniques. Rather, bootstrap techniques avoid the stringent parametric assumptions by creating an empirical distribution of the statistic(s) of interest, and, from this empirical distribution, the observed quantiles can be used to find confidence limits for the statistic of interest (e.g., Efron & Tibshirani, 1993).

When forming confidence intervals for ω, consideration of bootstrap techniques is especially important because the assumption of multivariate normality will likely be violated in many situations. In particular, the multivariate normality refers to the set of J items being multivariate normally distributed. However, the items of many measurement instruments are based on Likert scalings, possibly with relatively few response categories. In such situations the multivariate normality assumption of the set of items being multivariate normally distributed will tend to be violated and the aforementioned analytic procedure may not yield the nominal confidence interval coverage. Additionally, issues of ceiling and floor effects for some items often arise in applied research. A similar rationale in the context of coefficient α motivating the bootstrap technique is given in Yuan, Guarnaccia, and Hayslip (2003), where the bootstrap procedure is compared and ultimately recommended to the analytic confidence interval approach.

The two most common bootstrap techniques for confidence interval estimation are the percentile method and the bias-corrected and accelerated (BC*a*) method. The BC*a* method is the generally preferred implementation of the bootstrap, but the BC*a* depends on the percentile method so the discussion necessarily begins with the percentile method. The next two sections describe each of these methods as they apply to the estimation of confidence intervals for the population value of ω.

The Percentile Method

Suppose a random sample of N independent individuals each respond to all items on a measurement instrument with J items. The idea of bootstrapping is to sample, with replacement, the results of N measurement instruments B times, where B is the number of bootstrap replications and should be relatively large (e.g., $B = 10{,}000$). It is important to remember that in the context of the bootstrap, each individual's set of scores has an equal probability of being selected on each random sampling from the complete set of observed data. Specifically, each individual's set of scores has a probability of $1/N$ of being selected on any given random selection. These repeated samplings occur N times for each of the B bootstrap replications.

The idea of the percentile method is that the statistic of interest is calculated for each of the B bootstrap replications. These B statistics then form an empirical distribution (i.e., one not based on assumptions but on the observed distribution of the statistic of interest), where the percentiles of the empirical distribution are used as confidence limits for the population parameter of interest.

Let $\widehat{\boldsymbol{\omega}}^*$ be a vector of length B of each of the bootstrap estimates of ω. Suppose one is interested in the $(1 - \alpha')$ 100% confidence interval for ω, where α' designates the Type I error rate. The lower and upper confidence limits for a $(1 - \alpha')100\%$ symmetric confidence interval are defined as the value of $\widehat{\boldsymbol{\omega}}^*$ corresponding to the $\alpha'/2$ percentile and the $(1 - \alpha'/2)$ percentile, respectively.[8] Formally, the lower and upper confidence interval limits for the percentile method are given as

$$L_{\mathrm{PM}} = \widehat{G}^{-1}(\widehat{\boldsymbol{\omega}}^*|\alpha'/2) \tag{23}$$

and

$$U_{\mathrm{PM}} = \widehat{G}^{-1}(\widehat{\boldsymbol{\omega}}^*|1 - \alpha'/2), \tag{24}$$

where L_{PM} and U_{PM} are the lower and upper confidence limits of the percentile method, respectively, $\widehat{G}$ is the estimated cumulative distribution function of the distribution of values identified on the left of the given sign (|) at the specified quantile on the right of the given sign, and $\widehat{G}^{-1}$ is its inverse. By definition, $\widehat{G}^{-1}(\widehat{\boldsymbol{\omega}}^*|\alpha'/2)$ equals the $\alpha'/2$ quantile of $\widehat{\boldsymbol{\omega}}^*$. For example, if α' is .05 for a symmetric confidence interval, $\widehat{G}^{-1}(\widehat{\boldsymbol{\omega}}^*|.025)$ is the .025 quantile and $\widehat{G}^{-1}(\widehat{\boldsymbol{\omega}}^*|.975)$ is the .975 quantile from the bootstrap distribution of $\widehat{\boldsymbol{\omega}}^*$.

The Bias-Corrected and Accelerated Method

Conceptually, the BC*a* method is analogous to the percentile method in the sense that, from the empirical distribution of the statistic of interest (here $\widehat{\boldsymbol{\omega}}^*$), quantiles are found that represent the confidence limits of the population parameter of interest. The two methods of confidence interval construction differ in the particular quantile that is selected to form the limits of the confidence interval. Whereas the confidence limits for the percentile method are $\alpha'/2$ and $1 - \alpha'/2$, the confidence limits for the BC*a* method are dependent on these two and two other values obtained from the empirical distribution of the statistic of interest, which we now briefly discuss.

The bias-correction estimate, denoted $\widehat{z}_0$, is obtained by first determining the proportion of bootstrap replications less than the original estimate (Efron, 1998) and then finding the inverse of a standard normal distribution for that proportion.

That is,

$$\widehat{z}_0 = \Phi^{-1}\left(\frac{\sharp(\widehat{\boldsymbol{\omega}}^* < \widehat{\omega})}{B}\right), \tag{25}$$

where $\sharp$ represents "the number of" and Φ is the standard normal cumulative distribution function and Φ^{-1} its inverse (e.g., $\Phi(1.96) = .975$ and $\Phi^{-1}(.975) = 1.96$). When the distribution of $\widehat{\boldsymbol{\omega}}^*$ is perfectly symmetric, $\widehat{z}_0 = 0$.

The acceleration estimate $(\widehat{a})$, on the other hand, quantifies the rate of change of the standard error of the estimate, with respect to the true value of the parameter and is measured on a normalized scale (Efron, 1998, 1987). Unfortunately, an intuitive explanation for $\widehat{a}$ is not readily available, as derivation of its estimator depends on a higher level of statistical knowledge than assumed in the present article. Nevertheless, the definition for $\widehat{a}$ can be given which may help in the understanding of what a computer program does when it computes $\widehat{a}$. Calculation of $\widehat{a}$ depends on the jackknife estimation procedure, which estimates $\widehat{\omega}$ with each of the *N* observations removed one-by-one (see, e.g., Miller, 1974 for a review). Let $\widehat{\omega}_{(-i)}$ be the value of $\widehat{\omega}$ when the *i*th data point has been deleted from the original sample and $\tilde{\omega}$ be the mean of the *N* jackknife $\widehat{\omega}_{(-i)}$ values. The acceleration is then computed as

$$\widehat{a} = \frac{\sum_{i=1}^{N}\left(\tilde{\omega} - \widehat{\omega}_{(-i)}\right)^3}{6\left[\sum_{i=1}^{N}\left(\tilde{\omega} - \widehat{\omega}_{(-i)}\right)^2\right]^{\frac{3}{2}}}. \tag{26}$$

Thus, the BC*a* is a more accurate method of confidence interval formation as compared to the percentile method.

Given the characteristics of the empirical distributions, in particular the bias and acceleration of the rate of change of the standard error of the estimate with respect to the population value, the BC*a* method increases the accuracy of the obtained confidence interval. In fact, the BC*a* method is second-order accurate, whereas the percentile method is only first-order accurate. This means that confidence intervals from the percentile method approach the correct value of confidence interval coverage at a rate of $1/\sqrt{N}$, whereas the BC*a* method approaches the correct value of the confidence interval coverage at a rate of 1/*N* (Efron & Tibshirani, 1993, pp. 187–188).

Given $\widehat{z}_0$ and $\widehat{a}$, the lower and upper confidence limits of BC*a* confidence intervals are obtained by

$$L_{\mathrm{BC}a} = \widehat{G}^{-1}\left(\widehat{\boldsymbol{\omega}}^*|\Phi\left(\widehat{z}_0 + \frac{\widehat{z}_0 + (z_{\alpha'/2})}{1 - \widehat{a}\left(\widehat{z}_0 + z_{(\alpha'/2)}\right)}\right)\right) \tag{27}$$

and

$$U_{\mathrm{BC}a} = \widehat{G}^{-1}\left(\widehat{\boldsymbol{\omega}}^*|\Phi\left(\widehat{z}_0 + \frac{\widehat{z}_0 + z_{(1-\alpha'/2)}}{1 - \widehat{a}\left(\widehat{z}_0 + z_{(1-\alpha'/2)}\right)}\right)\right) \tag{28}$$

respectively, where $L_{\mathrm{BC}a}$ and $U_{\mathrm{BC}a}$ are the lower and upper confidence limits of the BC*a* method, respectively. Notice that when $\widehat{a}$ and $\widehat{z}_0$ equal zero, $L_{\mathrm{BC}a} = \widehat{G}^{-1}(\widehat{\boldsymbol{\omega}}^*|\alpha'/2)$ and $U_{\mathrm{BC}a} = \widehat{G}^{-1}(\widehat{\boldsymbol{\omega}}^*|1 - \alpha'/2)$, which are equal to the confidence limits of the percentile method.

Although the percentile and BC*a* confidence interval methods are statistically and conceptually appealing, such confidence intervals are computationally intensive. Bootstrap confidence intervals require randomly selecting *B* random bootstrap data sets, fitting a confirmatory factor model with the appropriate constraints, collecting the output of *B* bootstrap replications, and implementing the methods outlined for the bootstrap confidence interval procedure.

[8] Note that the confidence interval need not be symmetric. For example, one could form a 95% confidence interval where there were four percentage points of the Type I error rate on the lower side of the distribution and only one percentage point of the Type I error rate on the upper side of the distribution.

Table 1. Point estimate of ω and confidence interval limits for the cooperation, advocate/influence, and negotiation subscales

		Lower limit	Estimate	Upper limit
	Analytical	0.8542		0.9141
Cooperation	Percentile method	0.8466	.8841	0.9128
	BC*a*	0.8493		0.9147
	Analytical	0.7474		0.8526
Advocate/Influence	Percentile method	0.7384	.8000	0.8460
	BC*a*	0.7374		0.8453
	Analytical	0.7221		0.8371
Negotiation	Percentile method	0.7098	.7796	0.8332
	BC*a*	0.7137		0.8349

Note. BC*a* represents the bias-corrected and accelerated bootstrap procedure. The point estimate is the same for each of the three (analytic, percentile method, and BC*a*) methods of confidence interval formation.

The MBESS (Kelley & Lai, 2010; Kelley, 2007a, 2007b) R (R Development Core Team, 2010) package automates this task and implements bootstrap confidence intervals for ω with a simple to use function, which is illustrated in detail in Appendix B. In the next section, an example is shown to demonstrate the utility of the three methods of confidence interval construction for ω.

Empirical Example

Teams are an integral part of many organizations (e.g., Guzzo & Dickson, 1996; Kozlowski & Ilgen, 2006; Sundstrom, 1999). Objective measurements of the various dimensions of team effectiveness, collaboration, function, cohesion, etc., are important topics in many areas of applied research. The Ford Motor Company Partnership for Advanced Studies (Ford PAS) (Zhuang, MacCann, Wang, Liu, & Roberts, 2008) is a program that "provides students with content knowledge and skills necessary for future success – in such areas as business, economics, engineering, and technology" (Ford Motor Company Fund, 2008–2010). In order to assess various dimensions of the Ford PAS program, 159 participants from high school (77 male and 82 female; mean/standard deviation of age was 16.10/1.03; 64.2% African-American, 18.9% White non-Hispanic, 3.1% Hispanic, 3.1% multiethnic, and 10.7% Native American, Asian, or Other) responded to a self-report questionnaire (see Zhuang et al., 2008; Wang, MacCann, Zhuang, Liu, & Roberts, 2008 for details). The questionnaire consisted of 30 items on a 6-point Likert scale with lower anchor 1 (never) and upper anchor 6 (always).[9] The questionnaire was used to measure three constructs: (a) Cooperation (12 items), (b) Advocate/Influence (9 items), and (c) Negotiation (9 items) (Zhuang et al., 2008). Performing listwise deletion on the 30 items of interest where missing data occurred yielded a sample size of 127.

The estimated composite reliability, assuming a congeneric factor structure (i.e., where ω is most appropriate as coefficient α's assumption of essential τ-equivilance would likely be violated), of each of the three subscales, assuming a congeneric factor structure, is .8841 for the Cooperation subscale, .8000 for the Advocate/Influence subscale, and .7796 for the Negotiation subscale, respectively. Of course, the point estimates themselves are fallible and interest does not literally revolve around the sample value, but rather the population value. Correspondingly, confidence intervals for the population reliability coefficients in each of these situations are needed. The point estimates themselves as well as the confidence interval limits are available from MBESS, where the way in which they are calculated is illustrated in Appendix B.

Table 1 displays the confidence intervals coming from the three different methods for the ω coefficient of the Negotiation subscale. As can be seen from Table 1, for each of the subscales the estimated confidence interval limits for the three methods in Table 1 tend to be fairly close to each other. One reason why this may occur is because in this situation the sample size may be sufficiently large and/or multivariate normality may hold approximately. Furthermore, Yuan and Bentler (2002) showed that the asymptotic confidence interval methods can be quite robust to violations of the multivariate normality assumption under a variety of conditions.[10]

With regard to which method of confidence interval construction should be used in general, we take a moderate approach suggested by Kelley (2005), which is a

[9] Originally, there were 57 items evaluated for inclusion in the questionnaire. Of the original 57, 27 items were eliminated based on psychometric principles.

[10] Because we do not advocate for α, we did not compute it here nor have we discussed formally in Appendix B. However, we should note that (Green & Yang, 2009a) showed that the true reliability and that estimated by coefficient α can be quite different. In the examples Green and Yang (2009a) showed, where the loadings for the general factor vary using 14 combinations of loadings of 0.20, 0.50, and 0.80 for a 6-item scale, the percentage of bias ranges from 0% to 11.10%. Green and Yang (2009a) also show how the bias decreases for similar situations for a 12-item scale, where the percentage of bias ranges from 0% to 5.1%. Correspondingly, we believe that coefficient ω should always be reported for scales designed to measure a single factor with uncorrelated errors.

compromise where both the analytic confidence interval and the bootstrap (ideally the BC*a*, we argue) confidence interval are reported. Such a compromise works well because some readers will not be familiar with the bootstrap approach and/or may believe the results were not as impressive as when using the analytic procedure (or they would have been presented since it is a parametric procedure and widely seen by some as preferable to a nonparametric procedure). However, ignoring the advancements and benefits of a bootstrap approach to statistical inference so as to satisfy some reviewers/editors/readers is also unfortunate. Correspondingly, presenting both results will presumably satisfy both types of readers and will allow readers to make more informed conclusions. Similarly, for those critical of the bootstrap approach, the results of the parametric alternative are provided. Thus, no single approach is given all of the weight when interpreting the results and when instances arise that lead to interpretational differences, the underlying data can be interrogated more thoroughly in an attempt to discover what might be contributing to major differences between the two approaches, as they should be similar if all assumptions are satisfied.

If only one confidence interval for the population reliability coefficient can be reported, the suggestion provided here is to use the BC*a* bootstrap approach, which is a general approach widely recommended in the methodological literature. When all assumptions of the model are realized, the three confidence interval approaches will tend to provide results that are similar. Given that the assumption of multivariate normality is not likely to hold for data often collected in social and behavioral settings (e.g., Micceri, 1989), the BC*a* approach is the choice we recommend.

Discussion

This article has served several purposes. First, it provided a review of the underlying fundamentals of composite scores. Second, it discussed the limitations of coefficient α, which is the most commonly used way to assess the reliability of a composite score in psychology and related disciplines. Third, the article discussed ω, which is a useful measure of reliability that is not well known and seldom used in the applied literature, even though its assumptions are more consistent with typical empirical data than the most widely used measure of reliability, coefficient α. Fourth, the article discussed methods of confidence interval formation (analytic and bootstrap) for reliability coefficients. Fifth and finally, the article develops software for the methods discussed and illustrates how the methods can easily be implemented with the freely available computer program R using the MBESS package. With the MBESS R package researchers can easily and immediately implement the methods discussed throughout the article. Taken together, the hope is that this article will be useful to researchers who use composite scores in their work and want to obtain statistically sound interpretations of the reliability of those composite scores.

Because of the importance of composite scores and their reliability in research, using the most appropriate estimate of the reliability is important. Furthermore, because of the importance of confidence intervals in modern research, and the unequivocal call for them to be reported (e.g., American Psychological Association, 2010; Grissom & Kim, 2005; Harlow, Mulaik, & Steiger, 1997; Hunter & Schmidt, 2004; Schmidt, 1996; Task Force on Reporting of Research Methods in AERA Publications, 2006; Thompson, 2002; Wilkinson & the American Psychological Association Task Force on Statistical Inference, 1999, etc.), the most meaningful confidence interval method should be used. Due to the assumption of multivariate normality likely being violated in many situations, the bootstrap approach is recommended, with the preferred type being the BC*a*. Even if a researcher appreciates the benefits of ω and the bootstrap confidence interval, implementation of those methods is not generally straightforward. However, with the MBESS R package, the intricate procedures discussed can easily be implemented, making the methods discussed in the present article immediately available to researchers.

References

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

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

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Duxbury Press.

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.

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

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

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of American Statistical Association, 82*, 171–185.

Efron, B. (1998). R. A. Fisher in the 21st century. *Statistical Science, 13*, 95–114.

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

Fan, X., & Thompson, B. (2001). Confidence intervals around score reliability coefficients, please: An EPM guidelines editorial. *Educational and Psychological Measurement, 61*, 517–532.

Ford Motor Company Fund (2008–2010). Retrieved from http://www.fordpas.org/

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.

Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach*. Mahwah, NJ: Erlbaum.

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.

Guzzo, R. A., & Dickson, M. W. (1996). Teams in organizations: Recent research on performance and effectiveness. *Annual Review of Psychology, 47*, 307–338.

Hahn, G., & Meeker, W. (1991). *Statistical intervals: A guide for practitioners*. New York, NY: Wiley.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

Jones, L. V., & Thissen, D. (2007). A history and overview of psychometrics. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (Vol. 26, pp. 1–27). New York, NY: Elsevier.

Jöreskog, K., & Sörbom, D. (1996). *LISREL 8: User's reference guide* (2nd ed.). Chicago, IL: Scientific Software International.

Kelley, K. (2005). The effects of nonnormal distributions on confidence intervals around the standardized mean difference: Bootstrapping as an alternative to parametric confidence intervals. *Educational and Psychological Measurement, 65*, 51–69.

Kelley, K. (2007a). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software, 20*(8), 1–24.

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

Kelley, K., & Lai, K (2010). MBESS 3.0 (or greater) [Computer software and manual]. Retrieved from http://www.cran.r-project.org/

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

Kozlowski, S. W. J., & Ilgen, D. R. (2006). Enhancing the effectiveness of work groups and teams. *Psychological Science in the Public Interest 7*, 77–124.

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

Li, H. (1997). A unifying expression for the maximal reliability of a linear composite. *Psychometrika, 62*, 245–249.

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

McArdle, J. J., & McDonald, R. P. (1984). Some algebraic properties of the reticular action model for moment structures. *British Journal of Mathematical and Statistical Psychology, 37*, 234–251.

McDonald, R. P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology, 38*, 1–21.

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

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156–166.

Miller, R. G. (1974). The jackknife – A review. *Biometrika, 61*, 1–15.

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.

Pawitan, Y. (2001). *In all likelihood: Statistical modelling and inference using likelihood*. New York, NY: Oxford University Press.

R Development Core Team. (2010). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria: R Foundation of Statistical Computing (ISBN 3-900051-07-0).

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

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

Revelle, W. (1979). Hierarchical cluster-analysis and the internal structure of tests. *Multivariate Behavioral Research, 14*, 57–74.

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

Schervish, M. J. (1995). *Theory of statistics*. New York, NY: Springer.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1*, 115–129.

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

Sijtsma, K. (2009b). Reliability beyond theory and into practice. *Psychometrika, 74*, 169–173.

Spearman, C. (1904). "General intelligence", objectively determined and measured. *American Journal of Psychology, 15*, 201–292.

Sundstrom, E. (1999). The challenges of supporting work team effectiveness. In E. Sundstrom (Ed.), *Supporting work team effectiveness* (pp. 3–23). San Francisco, CA: Jossey-Bass.

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.

Ten Berge, J. M. F. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika, 69*, 613–625.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*, 25–32.

Thompson, B. (2003). Understanding reliability and coefficient alpha, really. In B. Thompson (Ed.), *Score reliability: Contemporary thinking on reliability issues* (pp. 3–23). Thousand Oaks, CA: Sage.

Wang, L., Zhuang, X., Liu, L., MacCann, C., & Roberts, R. D. (2009). Assessing teamwork and collaboration in high school students: A multi-method approach. *Canadian Journal of School Psychology, 24*, 108–124.

Wilkinson, L., the American Psychological Association Task Force on Statistical Inference. (1999). Statistical methods in psychology: Guidelines and explanations. *American Psychologist, 54*, 594–604.

Yuan, K.-H., & Bentler, P. M. (2002). On robustness of the normal-theory based asymptotic distributions of three reliability coefficient estimates. *Psychometrika, 67*, 251–259.

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.

Zhuang, Z., MacCann, C., Wang, L., Liu, L., & Roberts, R. D. (2008). *Development and validity evidence supporting a teamwork and collaboration assessment for high school students*. ETS Research Report RR-08-50. Princeton, NJ: Educational Testing Service.

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., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and Mcdonald's ω_h: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*, 123–133.

Received June 23, 2009
Accepted November 22, 2010
Published online August 5, 2011

Ken Kelley

Department of Management
Mendoza College of Business
University of Notre Dame
Notre Dame, IN 46556
USA
Tel. +1 574 631-1459
Fax +1 574 631-5127
E-mail KKelley@ND.Edu, Web: http://www.nd.edu/∼ kkelley

Appendix A

In maximum likelihood theory, the first derivative of the likelihood function is termed the score function. The score function provides the maximum likelihood point estimates of the model parameters. The negative expected value of the derivative of the score function (i.e., the second derivative of the likelihood function) is the information matrix, often termed the Fisher information matrix. The inverse of the information matrix is the covariance matrix of the estimates. Thus, provided a likelihood function is twice differentiable and the appropriate assumptions met, the estimates (from the score function) and the variance/covariance of the estimates (from the inverse of the information matrix) can be obtained. Therefore, $Var(\widehat{\upsilon})$, $Var(\widehat{\nu})$, and $Cov(\widehat{\upsilon}, \widehat{\nu})$ from Equation 21 are readily available in a maximum likelihood context by way of the information matrix. A discussion of maximum likelihood estimation in general and the score function and information matrix in particular is provided by Pawitan (2001) and Schervish (1995).

In the present context, the sum of the values of the first J elements of the principal diagonal is the variance of $(\widehat{\upsilon})$ (i.e., the variance of the sum of the $J\,\widehat{\lambda}_j$s, $Var(\widehat{\upsilon})$) and the sum of the second set of J elements on the principal diagonal is the variance of $(\widehat{\nu})$ (i.e., the variance of the sum of the J ψ_j^2s, $Var(\widehat{\nu})$), with the sum of the off diagonal elements, the covariances of the $\widehat{\lambda}_j$s and ψ_j^2s, being the covariance of $(\widehat{\upsilon})$ and $(\widehat{\nu})$ (i.e., $Cov(\widehat{\upsilon}, \widehat{\nu})$).

To help solidify the above discussion, a general example is provided. For a measurement instrument with J items, the inverse of the information matrix, for parameter estimate set $\widehat{\boldsymbol{\theta}}$, can be partitioned as follows:

$$I\left(\widehat{\boldsymbol{\theta}}\right)^{-1} = \left[\begin{array}{c|c} Cov\left(\widehat{\lambda}\right) & Cov\left(\widehat{\lambda}, \widehat{\psi}^2\right) \\ \hline Cov\left(\widehat{\psi}^2, \widehat{\lambda}\right) & Cov\left(\widehat{\psi}^2\right) \end{array}\right],$$

where $I\left(\widehat{\boldsymbol{\theta}}\right)^{-1}$ is the inverse of the information matrix. Using the same ordering scheme as used in the Cartesian coordinate plane, quadrant I (top right) and III (bottom left) of $I\left(\widehat{\boldsymbol{\theta}}\right)^{-1}$ are equivalent and are denoted $Cov(\widehat{\lambda}, \widehat{\psi}^2)$. Quadrant II (top left) contains the variance/covariance matrix of the $\widehat{\lambda}$s, whereas quadrant IV (bottom right) contains the variance/covariance matrix of the $\widehat{\psi}^2$s.

As is shown in Equation 21, estimates of the variance of $\widehat{\upsilon}$, the variance of $\widehat{\nu}$, and the covariance of the two are necessary in order to estimate the standard error of ω. The variance of $\widehat{\upsilon}$ is the sum of *all* elements in quadrant II, that is, $Var(\widehat{u})$ is the sum of the variances of the $\widehat{\lambda}$s plus the sum of their covariances (i.e., the off diagonal elements in quadrant II). The variance of $\widehat{\nu}$ is the sum of *all* quadrant elements in IV, that is, $Var(\widehat{\nu})$ is the sum of the variances of the $\widehat{\psi}^2$s plus the sum of all off diagonal elements in quadrant IV. The covariance of the $\widehat{\lambda}$s and $\widehat{\psi}^2$s is then the sum of all elements in quadrant I or quadrant III (the two are equal because of their symmetry, as is always the case with a covariance matrix). Thus, $2Cov(\widehat{\upsilon}, \widehat{\nu})$ is simply the sum of all of the elements in quadrant I plus the sum of all of the elements in quadrant III (or twice the sum of either quadrant I or III). After $Var(\widehat{\upsilon})$, $Var(\widehat{\nu})$, and $Cov(\widehat{\upsilon}, \widehat{\nu})$ have been calculated, formation of the confidence interval for the reliability of an unweighted composite is straightforward via Equation 22. However, all of the methods discussed can be implemented with the MBESS package for the R program, which eliminates the need for dealing with the equations, estimation procedures, and computations directly, if one so desires.

Appendix B

The MBESS (Kelley, 2007a, 2007b; Kelley & Lai, 2010) package for the R program (R Development Core Team, 2010) package for the R program (R Development Core Team, 2010) can be used for the reported reliability estimates. The way to estimate the reliability for a composite variable of a set of scores with MBESS is by substituting the appropriate information into the `ci.reliability()` function as

```
ci.reliability(S=S, N=N, model=‘‘Congeneric’’,
type=‘‘Factor Analytic’’, conf.level=1 − α’),
```

where S is the observed covariance matrix of the items measuring a particular factor, N is the sample size, and $1-\alpha'$ is the desired level of confidence interval coverage.[11] The argument `model` in the `ci.reliability()` function is used to identify the model of interest, which can subsume either `Parallel`, `True-Score Equivalent` (i.e., coefficient α), or `Congeneric` (i.e., ω). The `type` argument can subsume either `Normal Theory` or `Factor Analytic` for the standard formula based approach or the factor analytic approach, respectively. Although any model can be estimated with the `type=‘‘Factor Analytic’’` specification, the `model=‘‘Congeneric’’` requires `type= ‘‘Factor Analytic’’`.

For example, the way `ci.reliability()` function can be used to estimate ω and the corresponding 95% confidence interval coverage for ω can be implemented as

```
ci.reliability(S=Cov.Negotiation,
N=127, model=‘‘Congeneric’’,
type=‘‘Factor Analytic’’, conf.level=.95),
```

where `Cov.Negotiation` is the covariance matrix of the nine items used to measure the negotiation factor. After submitting the above code, the function returns

```
$CI.lower
[1] 0.722051
$CI.upper
[1] 0.8370732
$Estimated.reliability
[1] 0.7795621
$SE.reliability
[1] 0.02934295
$Conf.Level
[1] 0.95.
```

As can be seen, the returned lower and upper confidence limits are .7221 and .8371, respectively, with the point estimate itself being .7796. What is important to realize is that the confidence interval implemented above is based on the analytic approach to confidence interval formation, which itself is quite involved as discussed in the analytic approach to confidence intervals section, but is easily implemented with the MBESS `ci.reliability()` function. Notice also that the standard error of the reliability coefficient is reported, as well as a reminder of the confidence interval coverage selected.

In addition to the analytic confidence intervals performed above, another MBESS function can be used to easily implement the bootstrap procedures discussed in the bootstrap sections. Now, bootstrap confidence intervals can be performed by using the `ci.reliability()` function with the option `Bootstrap=TRUE` specified, along with `B` (the number of bootstrap replications) and Bootstrap CI (type of bootstrap confidence interval). The `ci.reliability()` function for bootstrap BCa confidence intervals can be implemented as follows:

```
ci.reliability(data=Negotiation,
model=‘‘Congeneric’’
type=‘‘Factor Analytic’’, conf.level=
.95, Bootstrap = TRUE, B=10000, Bootstrap
CI=‘‘BCa’’)
```

where `data` is the full data set, `Negotiation` here, `B` is the desired number of bootstrap replications (a large number, such as 10,000, is recommended), with the other parameters being equivalent to those given in the `ci.reliability()` function. In this situation the `ci.reliability()` function returns the following:

```
             Desired. Lower.  Upper.    Lower.     Upper.
             Conf.    Limit.  Limit.    Conf.      Conf.
             Level    Index   Index     Limit      Limit
Percentile.  0.95     250.00  9750.00  0.7086157  0.8324301
Method
BCa. method  0.95     292.39  9791.26  0.7123922  0.8341273
```

For both the percentile and the BC*a* bootstrap methods an index is provided, which shows the relative position of the ordered bootstrap replicates. For the percentile method the values are the desired quantiles of the distribution (e.g., .025 and .975) multiplied by B, the number of bootstrap replications. Because the BC*a* method uses additional parameters to estimate the confidence interval limits, the indices will tend to differ from the well-defined indices used for the percentile method.

[11] Use of MBESS requires R and for MBESS to be installed within R and loaded. MBESS is loaded with the command `require (MBESS)` and can be installed directly from within R by using the package installation utility on the menu bar within Windows and Macintosh operating systems.