Journal of Modern Applied Statistical Methods

Volume 12 | Issue 1 Article 13

5-1-2013

Bootstrap Interval Estimation of Reliability via Coefficient Omega

Miguel A. Padilla
Old Dominion University, mapadill@odu.edu

Jasmin Divers
Wake Forest School of Medicine, Winston-Salem, NC

Journal of Modern Applied Statistical Methods
May 2013, Vol. 12, No. 1, 78-89

1538 – 9472/13/$95.00

Bootstrap Interval Estimation of Reliability via Coefficient Omega

Miguel A. Padilla
Old Dominion University
Norfolk, VA

Jasmin Divers
Wake Forest School of Medicine
Winston-Salem, NC

Three different bootstrap confidence intervals (CIs) for coefficient omega were investigated. The CIs were assessed through a simulation study with conditions not previously investigated. All methods performed well; however, the normal theory bootstrap (NTB) CI had the best performance because it had more consistent acceptable coverage under the simulation conditions investigated.

Key words: Coefficient omega, reliability, composite reliability, bootstrap, confidence interval, interval estimate.

Introduction

Coefficient omega was proposed in the literature over 40 years ago (McDonald, 1970) as a reliability measure of homogenous items from a measurement instrument. It indexes the consistency with which the items measure the underlying latent variable (or construct). Based on a factor analytic model, coefficient omega uses the item factor loading and uniqueness to estimate reliability. Therefore, coefficient omega can be viewed as a more intuitive measure of reliability compared to coefficient alpha. However, it is rarely used in practice for two reasons: (1) it is largely overshadowed by coefficient alpha (Cronbach, 1951; Guttman, 1945), and (2) its theoretical framework is narrow with a limited body of knowledge about its properties with respect to statistical inference.

One issue faced by behavioral/social science researchers is the presence of measurement error in data collected through multiple-item questionnaires, inventories and other measurement instruments. The most common estimator of reliability used in the behavioral/social sciences is coefficient alpha (Hogan, Benjamin & Brezinski, 2000), at times referred to as Cronbach's coefficient alpha or Cronbach's alpha (Peterson, 1994). With reliability coefficients such as alpha, behavioral/social science researchers are able to evaluate the reliability of their items to aid in the creation of reliable measurement instruments.

Coefficient alpha's dominance is a result of five features. First and most notably, it is relatively simple to calculate and is a common option in popular statistical packages such as SAS and SPSS. Second, coefficient alpha can be calculated after a single test administration as opposed to requiring at least two test administrations. Third, it can be computed for continuous, ordinal or binary items; this advantage is notable when working with binary items such as right/wrong, true/false, etc. Fourth, different types of interval estimates for coefficient alpha have been developed (Maydeu-Olivares, Coffman & Hartmann, 2007; Padilla, Divers & Newton, 2012; Romano, Kromrey & Hibbard, 2010; van Zyl, Neudecker & Nel, 2000; Yuan, Guarnaccia & Hayslip, 2003). Fifth, a lack of other options, as well as more than a half century of cited research gives the impression that coefficient alpha is the only viable estimate of reliability.

Miguel A. Padilla is an Assistant Professor of Quantitative Psychology in the Department of Psychology. His research interests are in applied statistics and psychometrics. Email him at: mapadill@odu.edu. Jasmin Divers is an Associate Professor of Biostatistics in the Department of Biostatistical Sciences. His research interests are in statistical genetics and biostatistics. Email him at: jdivers@wfubmc.edu.

Although coefficient alpha is an excellent estimator of internal consistency when used correctly, it is biased when items are not at least Tau-equivalent or essentially Tau-equivalent (Graham, 2006; Lord, Novick & Birnbaum, 1968; McDonald, 1999; Zinbarg, Revelle, Yovel & Li, 2005). Tau-equivalence of items can best be described within the framework of the classical true score model (CTSM) from classical test theory. The CTSM for items can be ordered from most to least restrictive as follows: (1) parallel, (2) Tau - or essentially Tau-equivalent, and (3) congeneric. For items under conditions 1 and 2, coefficient alpha is equal to the reliability of the set of items. Under condition 3, coefficient alpha underestimates the reliability for a set of items (Zinbarg, et al., 2005). However, coefficient omega is equal to the reliability of a set of items for all 3 conditions (McDonald, 1999; Zinbarg, et al., 2005).

As stated, a reason for the limited use of coefficient omega is the limited knowledge about its statistical properties. One noteworthy drawback is a lack of development and investigation of a confidence interval (CI) for coefficient omega.

Raykov (1998) proposed a bootstrap percentile CI for the composite reliability of congeneric items measuring a common dimension (Raykov, 1997). The method is specified as a structural equation model (SEM) and shows promise. An illustration of the method was applied to a small simulation that included a sample size of 400, 6 multivariate normal congeneric items, and assumed unidimensionality. The bootstrap estimates were based on 1,000 bootstrap samples.

In another study Raykov (2002) derived the standard error for composite reliability CIs via the delta method. As previously, the model was specified through an SEM framework and showed promise when illustrated with a small simulation (n = 500, 5 multivariate normal congeneric items) assuming unidimensionality. The delta method CI was also compared to the bootstrap percentile CI with 2,000 bootstrap samples; both methods had similar results.

In a parallel study, Raykov and Shrout (2002) presented a more general form of composite reliability in a SEM framework with bootstrapped percentile CIs. The method extends the previous method by Raykov (1997, 1998). The authors applied the method to a small simulation with a sample size of 300, 6 multivariate normal congeneric items, and assumed two dimensions. The study results show that the composite reliability estimate is unbiased and the CIs contain the population parameter. The bootstrap percentile CIs were based on 1,000 bootstrap samples.

More recently (Raykov, 2012; Raykov & Marcoulides, 2011), the non-bootstrap method has been illustrated using large example data sets (i.e., $n \geq 350$). Results indicate that the method is applicable to approximately continuous items having a multi-normal distribution. In addition, the method is applicable to non-normal items with at least 5 to 7 response categories with the use of the robust maximum likelihood estimator (MLR). Study results also indicate that the MLR estimator can be used with items with less than 5 response categories by using parcels. For further details about using the MLR estimator with parcels see Raykov and Marcoulides (2011).

Hence, studies of CIs for composite reliability have been conducted (Raykov, 1998, 2002), but there are several limitations. First, it is difficult to generalize findings based on small simulations or example data. Second, the simulation studies designed to test the proposed methods were based on continuous items; such items are rare in the behavioral/social sciences (Raykov, 2002), most are Likert/ordinal or, in some cases, binary. Third, the methods require large sample sizes based on the asymptotic theory that underlies SEMs – recall that maximum likelihood (ML) is the standard method of estimating SEMs (Bollen, 1989; Raykov, 1998, 2002; Raykov & Shrout, 2002). Lastly, the methods require specialized SEM software (e.g., EQS, Mplus, etc.).

This study assesses the performance of bootstrap CIs for composite reliability as specified through coefficient omega in terms of a one-factor model under simulation conditions that investigate the limitations above. Of particular interest is the impact of binary and Likert/ordinal (e.g., categorical) items, and a sample size less than 300 on the coefficient omega bootstrap CIs.

Coefficient Omega and Reliability

Consider a set of k items $x_1, x_2, ..., x_k$ designed to measure a single construct or attribute. A common procedure in behavioral/social science research is to estimate the reliability of the composite or sum score $x = \sum_{j=1}^{k} x_i$. This represents the reliability of the measurement instrument or test reliability. Reliability of the composite or sum score is defined as

$$\rho = \frac{\text{var}(\tau)}{\text{var}(x)} = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_u^2} \quad (1)$$

where var(.) denotes the variance operator, σ_τ^2 the true score variance, and σ_u^2 the error variance. This definition of reliability assumes that all items are parallel (Allen & Yen, 1979; Crocker & Algina, 1986).

The analogous composite reliability for congeneric items via coefficient omega is defined as

$$\omega = \frac{\left(\sum_{j=1}^{k} \lambda_j\right)^2}{\left(\sum_{j=1}^{k} \lambda_j\right)^2 + \sum_{j=1}^{k} \psi_j} \quad (2)$$

where λ_j and ψ_j are the j^{th} factor loading and uniqueness, respectively (McDonald, 1970, 1999). Note that this model assumes the items are measuring a single construct or factor (i.e., a one-factor model). Coefficient omega is estimated by replacing λ_j and ψ_j with the sample estimates $\hat{\lambda}_j$ and $\hat{\psi}_j$ in equation 2. Although several methods are available for estimating the factor loadings, ML will be used herein.

Bootstrapped Coefficient Omega CIs

The bootstrapping algorithm for coefficient omega can be summarized in three steps. Suppose $\mathbf{X} = \left(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\right)^t$ are the observed data where each $\mathbf{x}_i$ is a $1 \times k$ vector. First, obtain a bootstrap sample $\mathbf{X}^{(b)} = \left(\mathbf{x}_1^{(b)}, \mathbf{x}_2^{(b)}, ..., \mathbf{x}_n^{(b)}\right)^t$, which is the b^{th} random resample from $\mathbf{X}$ selected with replacement. Note that $\mathbf{X}$ and $\mathbf{X}^{(b)}$ have the same sample size. Second, compute the b^{th} bootstrap estimate of coefficient omega $\left(\hat{\omega}^{(b)}\right)$ from $\mathbf{X}^{(b)}$. Lastly, $\hat{\omega}^{(1)}, \hat{\omega}^{(2)}, \ldots, \hat{\omega}^{(B)}$ represents the empirical sampling distribution (ESD) for $\hat{\omega}$ for $b = 1, 2, \ldots, B$ bootstrap samples. The ESD can then be summarized for statistical inference about ω. Typical parameter estimates are the bootstrap mean, percentiles, quantiles and standard error (SE). The bootstrap estimate of SE is

$$SE\left(\hat{\omega}\right) = \left[\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\omega}^{(b)} - \bar{\omega}\right)^2\right]^{1/2} \quad (3)$$

where

$$\bar{\omega} = \frac{1}{B}\sum_{b=1}^{B}\hat{\omega}^{(b)} . \quad (4)$$

The three most common bootstrap CIs were examined. First, the normal theory bootstrap (NTB) CI is computed as $\hat{\omega} \pm Z_{\alpha/2} SE\left(\hat{\omega}\right)$. Second, the percentile based (PB) CI is obtained by computing the $\alpha/2$ and $1-\alpha/2$ percentiles from the $\hat{\omega}$ ESD where α is the type I error rate. Third, the bias-corrected and accelerated (BCa) CI is an improved version of the PB CI in that it adjusts the PB CI $\alpha/2$ and $1-\alpha/2$ percentiles in two ways: (1) it makes a correction for bias, and (2) a correction for skewness (or acceleration). Note that the NTB CI assumes that the ESD is normally distributed, whereas the PB and BCa make no assumption about the shape of the ESD. For technical and theoretical details concerning the three bootstrap CIs investigated see Efron and Tibshirani (1998).

Methodology

Simulation Design

Four different conditions were investigated in a 4 (number of items) × 3 (correlation type) × 5 (number of item response categories) × 4 (sample size) Monte Carlo simulation design. A total of 240 conditions were investigated. All simulated items were binary or Likert-type (ordinal) in order to mimic items commonly found in behavioral/social science research; none of the items were continuous. For each simulation condition, 1,000 replications were obtained.

Binary and Likert-type items were generated using the Maydeu-Olivares et al. (2007) method. In brief, the method is:

1. Select a $k \times k$ population correlation matrix **P**, where k is the number of items and a set of thresholds $\boldsymbol{\tau}$ for categorization so that resultant items have a predetermined skewness and kurtosis.

2. Generate an $n \times k$ multivariate data $\mathbf{X}^* \sim N(\mathbf{0}, \mathbf{P})$, where n is the sample size.

3. Categorize the generated data $\mathbf{X}^*$ with $\boldsymbol{\tau}$ into data **X**. Each variable x in **X** is categorized by the thresholds as follows: $x = m$ if $\tau_m < x^* < \tau_{m+1}$ for $m = 0,\ 1,\ \ldots,\ M-1$ where $\tau_0 = -\infty$ and $\tau_M = \infty$, and M is the number of categories.

4. Compute the *true population* coefficient omega (ω) according to **P** and the thresholds in $\boldsymbol{\tau}$. See Maydeu-Olivares et al. (2007) for details.

5. Estimate coefficient omega $(\hat{\omega})$ bootstrapped CIs from the categorized data **X**.

6. Determine if the bootstrapped CIs includes the population coefficient omega (ω).

The specific simulation conditions investigated are as follows.

Number of Items (k).

Past research on coefficient alpha has examined various numbers of items ranging from two to twenty (Duhachek & Iacobucci, 2004; Enders, 2003; Maydeu-Olivares et al. (2007). To make results consistent, the following number of items were selected for this study: $k = 5, 10, 15, 20$.

Correlation Type (ρ).

Three different item correlation structures for **P** were investigated. The first two correlation structures were from a parallel-item one-factor model with common loadings $\lambda = .55$ or $.705$. These two models generated compound symmetric item correlation structures with $\rho = .30$ or $.50$, respectively. The third correlation structure was generated from a congeneric item one-factor model with loadings of $\lambda = .3, .4, .5, .6, .7$. The third item correlation was unstructured, and the same as the one generated by Maydeu-Olivares et al. (2007), but modified for cases with multiples of 5 items (the original called for multiples of 7 items).

Item Response Categories (IRCs).

Five item response categories were investigated: 2, 2, 3, 5 and 7. None of the items were continuous. For each response category, the first category was set to 0. For example, for an item with seven categories, the first category was set to 0 and the last to 6. The data generated from the specified item correlation matrix above (**P**) were categorized with $\boldsymbol{\tau}$ using the same methodology as Maydeu-Olivares et al. (2007).

For the binary items, $\boldsymbol{\tau}$ was chosen so that the resultant categorized items had skewness = 0 and kurtosis = –2, and skewness = 0.41 and kurtosis = –1.83, respectively. The second condition was investigated by Maydeu-Olivares et al. (2007). For the Likert-type items, $\boldsymbol{\tau}$ was chosen so that the resultant categorized items had skewness = kurtosis = 0.

Sample Size (n). The following sample sizes were investigated: n = 50, 100, 150, 200. These are common sample sizes in behavioral/social science research. In addition, Duhachek and Iacobucci (2004) noted that going

beyond a sample size of 200 reaches a point of diminishing returns for reliability estimates.

In each simulation replication, coefficient omega was estimated along with its corresponding bootstrap CIs. In this study, the 100(1 – α)% CIs for coefficient omega were estimated from a total of 2,000 bootstrap samples, where $\alpha = .05$. Relative bias for coefficient omega was calculated as

$$\hat{\omega}_{\text{bias}} = \frac{\hat{\omega} - \omega}{\omega}. \quad (5)$$

CI coverage was assessed using Bradley's (1978) liberal criteria, which is defined as $1 - 1.5\alpha \leq 1 - \alpha^* \leq 1 - 0.5\alpha$ where α^* is the true Type I error probability. Coverage is defined as the proportion of estimated CIs that contain the true population coefficient omega. Therefore, acceptable coverage for $\alpha = .05$ is given by $[.925, .975]$.

Results

Point Estimate Bias

The estimate of bias was investigated because it can have a major impact on bootstrap CIs. However, tables with all combinations of the simulation conditions were inspected and no bias was observed. In fact, the largest bias observed was $\hat{\omega}_{\text{bias}} = .04$.

Confidence Interval Coverage

The NTB method had the best performance in terms of coverage. However, the major impact on the CIs was the number of items. Thus, results are presented in the context of number of items.

5 Items

Only the BCa method was impacted by 5 items (see Table 1). The PB method had acceptable coverage under all simulation conditions, however, the BCa method tended to be impacted when the sample size was 100 or less. In this case the BCa coverage probability was below the acceptable range. There were three instances where the BCa coverage probability was below the acceptable range when the sample size was 150 or more.

10 Items

The NTB and PB methods had unacceptable coverage in one instance each (see Table 2). In each case the unacceptable coverage occurred with a sample size of 50 and with a compound item correlation matrix with $\rho = 0.56$.

15 Items

In this situation, all methods had at least one instance of unacceptable coverage (see Table 3). The NTB method had unacceptable coverage for the unstructured item covariance matrix with a sample size of 50 and 2 and 3 item response categories. For the PB method, unacceptable coverage occurred in two instances with a sample size of 50 and 3 item response categories. For the BCa method, the unacceptable coverage occurred with a sample size of 150 and 5 item response categories.

20 Items

In this condition, only the NTB and PB methods were impacted (see Table 4). The NTB method had unacceptable coverage in two instances in the unstructured item covariance matrix with a sample size of 50 and 2 item response categories. Conversely, unacceptable coverage for the PB method occurred with 3 or more item response categories and with a sample size of 100 or less.

CI Coverage Bands

In Figure 1 the 95% CI coverage band is displayed for each method by number of items across all simulation conditions and shows the impact of five items on all methods. In particular, the BCa is most impacted by five items because it tended to be the furthest from acceptable coverage and has the most variance. Another noticeable feature is that the NTB method tended to have coverage bands that were slightly above 95%, whereas the PB and BCa methods tended to have coverage bands below 95%. The PB method appears to be the most conservative.

Table 1: 95% Coverage Probabilities for 5 Items

IRC	n	ρ = 0.30				ρ = 0.56				ρ = Unstructured			
		50	100	150	200	50	100	150	200	50	100	150	200
2[a]	NTB	.945	.932	.944	.951	.940	.948	.944	.941	.925	.935	.942	.934
	PB	.949	.941	.945	.954	.949	.956	.948	.944	.944	.939	.940	.947
	BCa	.928	**.919**	.934	.948	.948	.952	.951	.945	.953	.926	**.917**	**.920**
2[b]	NTB	.936	.940	.937	.949	.940	.950	.952	.940	.940	.936	.948	.949
	PB	.954	.944	.943	.954	.942	.948	.946	.945	.954	.946	.942	.955
	BCa	.940	.929	.939	.949	.942	.950	.950	.945	.941	.932	.928	.934
3	NTB	.936	.950	.953	.937	.953	.940	.940	.949	.943	.950	.945	.938
	PB	.935	.941	.963	.937	.939	.939	.939	.944	.958	.948	.939	.940
	BCa	**.912**	.929	.957	.929	.936	.938	.938	.947	.934	.926	**.923**	.928
5	NTB	.938	.940	.944	.932	.948	.943	.943	.945	.933	.940	.947	.950
	PB	.936	.947	.940	.942	.933	.940	.940	.943	.937	.937	.951	.956
	BCa	**.920**	.939	.938	.937	.932	.936	.936	.943	**.920**	**.914**	.932	.953
7	NTB	.940	.930	.946	.943	.941	.949	.949	.935	.937	.940	.956	.944
	PB	.942	.934	.935	.944	.935	.943	.943	.936	.949	.953	.954	.948
	BCa	.925	.926	.935	.938	.932	.946	.946	.934	.927	.938	.945	.944

Notes: For IRC = 2[a], skewness = 0 and kurtosis = –2; for IRC = 2[b] skewness = 0.41 and kurtosis = –1.83. Bold numbers indicate unacceptable coverage outside [0.925, 0.975]. NTB = normal theory bootstrap; PB = percentile bootstrap; BCa = biased-corrected and accelerated bootstrap. All methods based on 2,000 bootstrap samples.

Table 2: 95% Coverage Probabilities for 10 Items

IRC	n	ρ = 0.30				ρ = 0.56				ρ = Unstructured			
		50	100	150	200	50	100	150	200	50	100	150	200
2[a]	NTB	.970	.961	.957	.957	.958	.957	.940	.957	.974	.968	.967	.949
	PB	.954	.955	.946	.950	.945	.949	.936	.961	.954	.964	.958	.945
	BCa	.964	.956	.951	.955	.950	.957	.939	.960	.963	.971	.964	.944
2[b]	NTB	.970	.963	.951	.970	**.979**	.946	.945	.941	.968	.961	.952	.956
	PB	.953	.952	.944	.963	.970	.947	.942	.941	.957	.948	.955	.944
	BCa	.955	.954	.944	.964	.975	.952	.944	.944	.957	.964	.962	.947
3	NTB	.965	.955	.944	.950	.951	.957	.951	.954	.950	.964	.950	.957
	PB	.926	.937	.951	.938	.927	.944	.944	.948	.926	.943	.943	.951
	BCa	.942	.942	.949	.944	.938	.955	.946	.954	.930	.952	.947	.950
5	NTB	.961	.957	.946	.947	.960	.950	.949	.960	.966	.955	.950	.946
	PB	.941	.945	.944	.941	.944	.943	.942	.950	.936	.949	.947	.947
	BCa	.952	.953	.941	.944	.948	.941	.946	.953	.941	.945	.948	.945
7	NTB	.961	.956	.947	.943	.946	.942	.943	.952	.964	.951	.938	.945
	PB	.931	.951	.944	.938	**.923**	.928	.943	.945	.9630	.936	.933	.942
	BCa	.935	.950	.946	.937	.928	.933	.939	.947	.939	.941	.934	.942

Notes: For IRC = 2[a], skewness = 0 and kurtosis = –2; for IRC = 2[b] skewness = 0.41 and kurtosis = –1.83. Bold numbers indicate unacceptable coverage outside [0.925, 0.975]. NTB = normal theory bootstrap; PB = percentile bootstrap; BCa = biased-corrected and accelerated bootstrap. All methods based on 2,000 bootstrap samples.

Table 3: 95% Coverage Probabilities for 15 Items

IRC	n	ρ = 0.30				ρ = 0.56				ρ = Unstructured			
		50	100	150	200	50	100	150	200	50	100	150	200
2[a]	NTB	.968	.973	.954	.949	.964	.955	.937	.943	**.979**	.966	.952	.952
	PB	.956	.963	.956	.952	.950	.950	.938	.944	.968	.950	.945	.951
	BCa	.966	.966	.962	.956	.959	.949	.942	.948.	**.980**	.955	.950	.953
2[b]	NTB	.971	.965	.943	.955	.954	.949	.959	.944	.972	.961	.953	.943
	PB	.949	.944	.940	.956	.945	.944	.958	.941	.946	.951	.941	.944
	BCa	.965	.957	.947	.958	.947	.948	.959	.945	.961	.962	.943	.951
3	NTB	.970	.953	.949	.953	.951	.955	.952	.954	**.976**	.957	.949	.953
	PB	.934	.938	.945	.945	**.921**	.940	.944	.950	**.922**	.937	.947	.946
	BCa	.952	.946	.949	.949	.933	.945	.954	.952	.954	.939	.945	.951
5	NTB	.956	.947	.930	.950	.956	.950	.949	.941	.965	.948	.951	.946
	PB	.928	.941	.926	.952	.934	.943	.947	.947	.939	.950	.950	.945
	BCa	.934	.945	**.919**	.952	.941	.950	.952	.937	.949	.944	.950	.947
7	NTB	.974	.945	.955	.947	.939	.950	.946	.940	.960	.950	.958	.944
	PB	.937	.930	.939	.941	.929	.947	.947	.931	.928	.943	.950	.945
	BCa	.943	.931	.942	.938	.929	.951	.947	.933	.932	.943	.952	.945

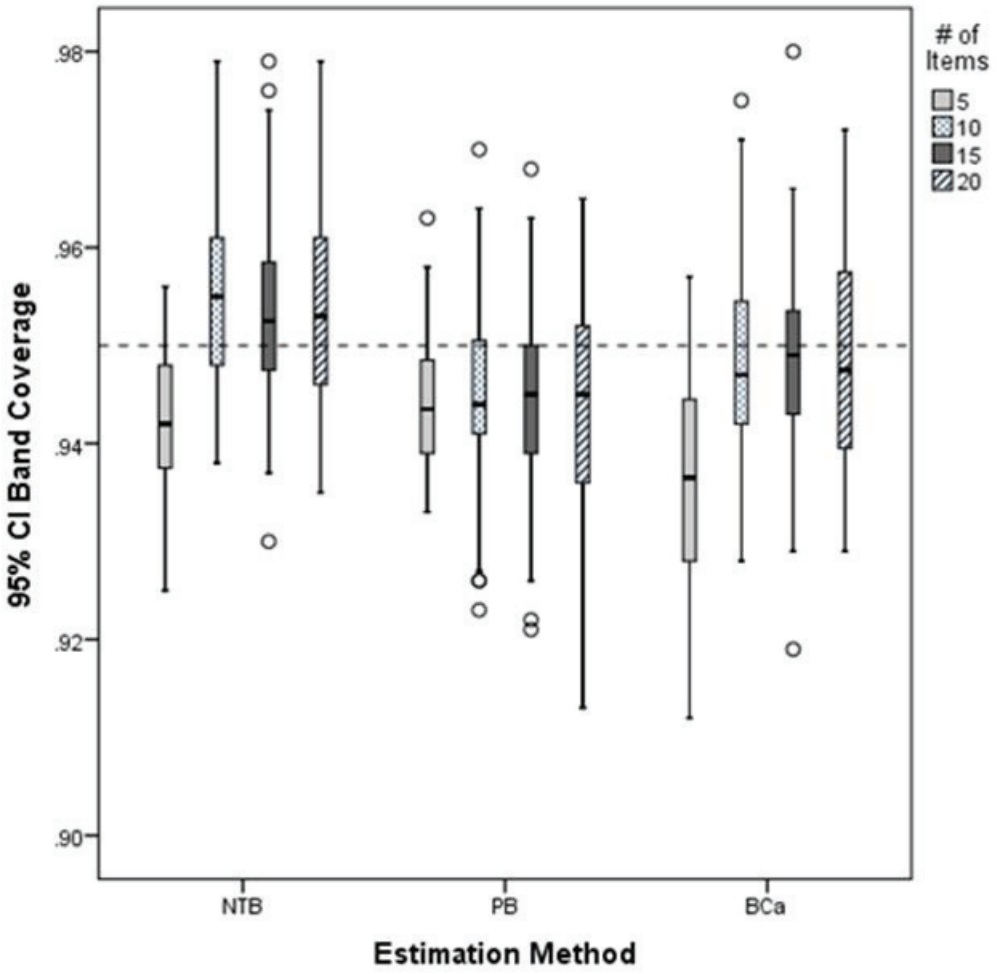
Notes: For IRC = 2[a], skewness = 0 and kurtosis = –2; for IRC = 2[b] skewness = 0.41 and kurtosis = –1.83. Bold numbers indicate unacceptable coverage outside [0.925, 0.975]. NTB = normal theory bootstrap; PB = percentile bootstrap; BCa = biased-corrected and accelerated bootstrap. All methods based on 2,000 bootstrap samples.

Table 4: 95% Coverage Probabilities for 20 Items

IRC	n	ρ = 0.30				ρ = 0.56				ρ = Unstructured			
		50	100	150	200	50	100	150	200	50	100	150	200
2^a	NTB	.971	.961	.943	.955	.965	.944	.935	.956	**.976**	.964	.948	.947
	PB	.953	.955	.940	.957	.955	.942	.941	.963	.940	.956	.945	.950
	BCa	.971	.960	.943	.960	.960	.949	.942	.961	.966	.966	.952	.953
2^b	NTB	.972	.969	.963	.965	.945	.947	.947	.951	**.979**	.958	.958	.958
	PB	.947	.964	.961	.965	.941	.949	.945	.944	.956	.943	.947	.955
	BCa	.968	.972	.966	.960	.943	.953	.949	.944	.965	.948	.951	.958
3	NTB	.971	.947	.956	.958	.948	.949	.958	.951	.973	.947	.947	.945
	PB	**.913**	.941	.951	.958	.927	.936	.953	.960	.947	.936	.943	.938
	BCa	.940	.942	.955	.957	.935	.939	.958	.956	.963	.935	.946	.942
5	NTB	.965	.936	.957	.949	.956	.948	.943	.949	.968	.959	.953	.945
	PB	.939	**.922**	.948	.949	**.924**	.934	.936	.946	**.923**	.952	.948	.946
	BCa	.940	.934	.947	.946	.933	.939	.941	.948	.932	.950	.950	.942
7	NTB	.961	.939	.944	.944	.953	.943	.956	.938	.966	.940	.939	.957
	PB	.929	.934	.940	.947	.928	.932	.949	.936	**.921**	.929	.939	.952
	BCa	.934	.934	.938	.947	.937	.939	.953	.944	.929	.931	.935	.952

Notes: For IRC = 2^a, skewness = 0 and kurtosis = –2; for IRC = 2^b skewness = 0.41 and kurtosis = –1.83. Bold numbers indicate unacceptable coverage outside [0.925, 0.975]. NTB = normal theory bootstrap; PB = percentile bootstrap; BCa = biased-corrected and accelerated bootstrap. All methods based on 2,000 bootstrap samples.

Figure 1: Distribution of 95% Bootstrap CI coverage for Estimation Method by Number of Items



Notes: CI = confidence interval; NTB = normal theory bootstrap; PB = percentile bootstrap; BCa = biased-corrected and accelerated bootstrap; dashed line is at 0.95.

Conclusion

Coefficient omega bootstrap CIs were proposed and their performance was investigated under several simulation conditions. Coefficient omega is a reliability index for a composite of congeneric items measuring a common dimension (i.e., factor or latent variable). To date, no study has investigated the performance of composite reliability CIs for congeneric items measuring a common dimension in such a simulation design. Results indicate that the NTB CI had the best coverage across all of the simulation conditions investigated. Even so, the major impact on the coefficient omega bootstrap CIs were the number of items.

Although the number of items impacted all three bootstrap CIs, it was most noticeable for the BCa method. In general, all methods were impacted when the number of items was set to five in that they tended to have coverage probability below 95% on average. However, only the BCa method had unacceptable coverage with five items. In fact, it had unacceptable coverage in eight instances, five of which had the unstructured covariance matrix. This suggests that unstructured covariance matrices

for five items do not provide enough information for the BCa method to make the proper adjustments.

When there are ten or more items, with the exception of the BCa method, the results were somewhat sporadic. In this case, the BCa method had acceptable coverage under all but two simulation conditions. The NTB method tended to have unacceptable coverage with fifteen or more items, a sample size of 50, and an unstructured item covariance matrix. Conversely, the PB method tended to have unacceptable coverage with a sample size of 50. For the NTB and PB method, most of the unacceptable coverage was so close to the $[.925, .975]$ boundaries that failure to fall within the interval was likely due to sampling variability.

Within the contexts of the simulation conditions investigated there is a clear order of preference of the bootstrap CIs investigated. The NTB method had the best performance in that it had consistent acceptable coverage under all but five simulation conditions (235/240 = 0.979). This was followed by the PB and BCa methods, whose performances were comparable (232/240 = 0.967 for PB vs. 230/240 = 0.958 for BCa). Another noticeable feature was that the NTB method tended to be the most liberal and the PB method most conservative. Nevertheless, a recommendation can be made. When there are 10 or less items the NTB method performed well, however, when there are 15 or more items, the BCa method was superior. In light of these findings, it is important to emphasize that all three methods had an acceptable range of coverage within the context of the investigated simulation conditions.

Despite these promising results, more research is needed. These results were obtained assuming that the Likert/ordinal items were normally distributed or that the underlying distribution did not depart greatly from normality. However, it is unlikely that data will follow a normal distribution in applied settings. Therefore, future research should focus on the CI estimation of coefficient omega using data that deviate from normality.

Through the simulation results provided and because coefficient omega is a general index of reliability, five advantages can be pointed out about its corresponding bootstrap CIs. First, the investigated items were not continuous and this had no significant impact on the CIs (recall that all items investigated were binary or Likert/ordinal with response categories that ranged from 2 to 7). Second, a sample size of 50 to 200 did not have a major impact. This is a significant finding because the factor loadings that are used by coefficient omega are estimated through ML which is based on the law of large numbers. Therefore, the literature has noted that this is a condition in need of investigation (Raykov, 1998, 2002; Raykov & Shrout, 2002). Only the PB method appeared to be somewhat affected by a small sample size. Third, the type of correlation structure did not have a major impact, thus, coefficient omega appears to be appropriate for items that range from parallel to congeneric. Fourth, though not investigated in this study, coefficient omega can be used with measures that have multiple factors or latent variables (McDonald, 1970, 1999). Lastly, the methods investigated do not require specialized SEM software; they only require the freely available and general R statistical package (http://www.r-project.org/). As such, interested researchers can obtain an easy-to-use R function for the coefficient omega bootstrap CIs with example data free of charge by visiting the corresponding author's website (http://www.omegalab-padilla.org/).

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole Pub. Co.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(*2*), 144-152.

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Holt, Rinehart, and Winston.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(*3*), 297-334.

Duhachek, A., & Iacobucci, D. (2004). Alpha's standard error (ASE): An accurate and precise confidence interval estimate. *Journal of Applied Psychology*, *89*(*5*), 792-808.

Efron, B., & Tibshirani, R. (1998). *An introduction to the bootstrap.* Boca Raton, FL: CRC Press.

Enders, C. K. (2003). Using the expectation maximization algorithm to estimate coefficient alpha for scales with item-level missing data. *Psychological Methods*, *8*(*3*), 322-337.

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement*, *66*(*6*), 930-944.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*(*4*), 255-282.

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement*, *60*(*4*), 523-531.

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

Maydeu-Olivares, A., Coffman, D. L., & Hartmann, W. M. (2007). Asymptotically distribution-free (ADF) interval estimation of coefficient alpha. *Psychological Methods*, *12*(*2*), 157-176.

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*, *23*(*1*), 1-21.

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

Padilla, M. A., Divers, J., & Newton, M. (2012). Coefficient Alpha bootstrap confidence interval under nonnormality. *Applied Psychological Measurement*, *36*(*5*), 331-348.

Peterson, R. A. (1994). A Meta-Analysis of Cronbach's Coefficient Alpha. *Journal of Consumer Research*, *21*(*2*), 381-391.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, *21*(*2*), 173-184.

Raykov, T. (1998). A method for obtaining standard errors and confidence intervals of composite reliability for congeneric items. *Applied Psychological Measurement*, *22*(*4*), 369-374.

Raykov, T. (2002). Analytic estimation of standard error and confidence interval for scale reliability. *Multivariate Behavioral Research*, *37*(*1*), 89-103.

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

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

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

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*(*3*), 376-393.

van Zyl, J., Neudecker, H., & Nel, D. (2000). On the distribution of the maximum likelihood estimator of Cronbach's alpha. *Psychometrika*, *65*(*3*), 271-280.

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*(*1*), 5-23.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β and Mcdonald's ω_H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(*1*), 123-133.