[Journal of Modern Applied Statistical Methods](#)

Volume 16 | Issue 2 Article 9

December 2017

Performance Evaluation of Confidence Intervals for Ordinal Coefficient Alpha

Heather J. Turner
Dallas Independent School District, heatherjeanturner@gmail.com

Prathiba Natesan
University of North Texas, prathiba.natesan@unt.edu

Robin K. Henson
University of North Texas, robin.henson@unt.edu

Recommended Citation
Turner, H. J., Natesan, P., & Henson, R. K. (2017). Performance Evaluation of Confidence Intervals for Ordinal Coefficient Alpha. Journal of Modern Applied Statistical Methods, 16(2), 157-185. doi: 10.22237/jmasm/1509494940

Journal of Modern Applied Statistical Methods
November 2017, Vol. 16, No. 2, 157-185.
doi: 10.22237/jmasm/1509494940

ISSN 1538 − 9472

Performance Evaluation of Confidence Intervals for Ordinal Coefficient Alpha

Heather J. Turner
Dallas ISD
Dallas, TX

Prathiba Natesan
University of North Texas
Denton, TX

Robin K. Henson
University of North Texas
Denton, TX

The aim of this study was to investigate the performance of the Fisher, Feldt, Bonner, and Hakstian and Whalen (HW) confidence intervals methods for the non-parametric reliability estimate, ordinal alpha. All methods yielded unacceptably low coverage rates and potentially increased Type-I error rates.

Keywords: Ordinal alpha, confidence interval, coverage, Monte Carlo, simulation

Introduction

Reliability is an estimate of the consistency of results from a measurement (Crocker & Algina, 2008; Cronbach, 1951) and an essential component to establish validity of a scale (Allen & Yen, 1979, 2002). Social scientists often measure attitudes and opinions with ordinal Likert-type ratings. The individual options on the scale are assumed to be discrete realizations of an underlying continuously-scaled construct (Flora & Curran, 2004). Nevertheless, researchers often treat ordinally-scaled data as continuous by using statistical methods that assume continuity of data. This causes an empirical mismatch with the data analyzed (Gadermann, Guhn, & Zumbo, 2012; Schmitt, 1996; Sijtsma, 2009; Streiner, 2003), underestimation of sample coefficient alpha, and may lead researchers to incorrect conclusions (Duhachek & Iacobucci, 2004; Flora & Curran, 2004; Gadermann et al., 2012; Maydeu-Olivares, Coffman, & Hartmann, 2007; Zumbo, Gadermann, & Zeisser, 2007).

Dr. Turner is an evaluation analyst in the Dallas Independent School District. Email her at heatherjeanturner@gmail.com. Dr. Natesan is an Associate Professor of Research Methods, Measurement, and Statistics. Email her at prathibachaj@gmail.com. The first and second authors contributed equally to this manuscript and are listed alphabetically by first name. Dr. Henson is a Professor of Educational Psychology. Email him at robin.henson@unt.edu.

One strategy to overcome the violation of continuity is to use ordinal coefficient alpha, which uses polychoric correlations instead of Pearson correlations (Gadermann et al., 2012; Zumbo et al., 2007). Although ordinal coefficient alpha has been shown to be a more appropriate measure of reliability for ordinal data, it is still just a point estimate. Fan and Thompson (2001) emphasized the need to report confidence intervals for coefficient alpha as a method for assessing the relative stability of the statistic as opposed to invoking rigid cutoff decisions about whether a value is large enough or not. For instance, an alpha coefficient value of 0.8 is generally considered acceptable (Cronbach & Shavelson, 2004). A 95% confidence interval is expected to contain the true value of the statistical estimate 95% of the time when resampled. This resampling is often hypothetical. Confidence intervals are a function of the standard error of the statistic and their coverage rates indicate Type-I error rate. The probability of Type-I error increases when the coverage rate of confidence intervals is less than expected. Therefore, it is important to examine the performance of confidence intervals for statistical estimates. One such diagnostic is coverage rate. Coverage rate is defined as the percentage of confidence intervals that contain the true value. In empirical research we do not know what the true value is. However, in simulation the true value is known. Comparing the coverage rate to confidence level through simulation helps us verify if the confidence interval given by the theoretical formulas are accurate.

There are several confidence interval approaches for conventional coefficient alpha (e.g. Bonett, 2002; Feldt, 1965; Fisher, 1950; Hakstian & Whalen, 1976). Various confidence interval methods were investigated for conventional coefficient alpha (Bonett, 2002; Cui & Li, 2012; Duhachek & Iacobucci, 2004; Feldt, 1965; Fisher, 1950; Hakstian & Whalen, 1976; Maydeu-Olivares et al., 2007; Padilla, Divers, & Newton, 2012; van Zyl, Neudecker, & Nel, 2000; Yuan & Bentler, 2002). However, the confidence intervals have not been investigated for ordinal coefficient alpha. Given the widespread use of Likert-type data in educational and behavioral research and the increased emphasis on reporting and interpreting confidence intervals of estimates (Cohen, 1994; Cumming, 2012; Cumming & Fidler, 2009; Finch, Cumming, & Thomason, 2001; Thompson, 2006a, 2006b; Wilkinson & APA Task Force on Statistical Inference, 1999), there is a need to evaluate the performance of currently available confidence interval methods for the ordinal coefficient alpha.

The purpose of this study is to investigate the coverage rates, widths, and biases of the four types of confidence intervals (Bonnet, Feldt, Fisher, and Hakstian Whalen), and the accuracy of the ordinal alpha point estimates under varying data conditions. Sample size, the number of items on an instrument, skewness of the responses, and population alpha were chosen as the data conditions because these have known to have an impact on the confidence intervals of alpha (Cui & Li, 2012; Duhachek & Iacobucci, 2004; Romano, Kromrey, Owens, & Scott, 2011).

Literature Review

The reliability of a composite score may be estimated in a factor model as the ratio of item variances to total variances. The factor analytic representation of classical test theory is expressed as

$$X_i = \lambda_i \xi + u_i \quad i = 1, 2, \ldots, p, \tag{1}$$

where X_i denotes the observed scores on the i^{th} item, λ_i denotes the factor pattern coefficient of the ith item, ξ is the true score common factor, and u_i is the uniqueness or random error up to p number of items. Novick and Lewis (1967) derived coefficient alpha as an unbiased estimate when the factor coefficients of each variable are equal to the common factor. Coefficient alpha can be calculated in a factor model as

$$\rho_{xx} = \frac{\left(\sum_{i=1}^{p} \lambda_i\right)^2}{\left(\sum_{i=1}^{p} \lambda_i\right)^2 + \sum_{i=1}^{p} \operatorname{var}(e)_{ii}} \tag{2}$$

where $\operatorname{var}(e)_{ii}$ is the error variance of the i^{th} item in a factor analytic model. Generally score reliability increases as coefficient alpha approaches a value of one.

Ordinal Coefficient Alpha

Gadermann et al. (2012) recommended using a non-parametric estimate of reliability coefficients for ordinal data, especially when there are few item response categories and skewed response distributions. Ordinal alpha is analogous to coefficient alpha, only differing by the type of correlation matrix used for computation. While coefficient alpha uses the Pearson correlation matrix and assumes data are continuously scaled, ordinal alpha uses the polychoric correlation matrix.

Polychoric correlation estimates the linear relationship between two ordinally scaled variables as the correlation between their respective underlying latent variable distributions (Jöreskog, 1990). By not assuming normality for the underlying distribution, the thresholds are allowed to be unequally spaced. The observed ordinal response y_j for item j with C response categories, where the response option $c = 0, 1, 2, \ldots, C - 1$ is defined as

$$y_j = c, \text{ if } \tau_c < y_j^* < \tau_{c+1}, \tag{3}$$

where τ_c, τ_{c+1} are the thresholds on the underlying continuum y_j^* and satisfy the constraint

$$-\infty = \tau_0 < \tau_1 < \cdots < \tau_{C-1} < \tau_C = \infty. \tag{4}$$

The polychoric correlation, Φ, between two ordinal items y_i and y_j is given by the Pearson product-moment correlation between their corresponding underlying latent variables y_i^* and y_j^*, respectively. By treating the observed item's responses in this manner, ordinal alpha is a nonparametric reliability estimate. The formula for ordinal alpha is expressed as

$$a_{ordinal} = \left[\frac{k}{k-1}\right] * \left[\frac{k\left(\lambda^2\right) - h^2}{k\left(\lambda^2\right) + \mu^2}\right], \tag{5}$$

where $\alpha_{ordinal}$ is ordinal alpha, k is the number of items, λ^2 is the squared factor pattern coefficient, h^2 is the communality where for a 1-factor model $h^2 = \lambda^2$, μ is uniqueness ($\mu^2 = 1 - h^2$). Theoretically, ordinal alpha equals the true reliability when the items are tau-equivalent and fit a one-factor model with equal factor pattern coefficients (Maydeu-Olivares et al., 2007; Zumbo et al., 2007).

Confidence Intervals

Authors should report a reliability coefficient even when the focus is not psychometric because it is a critical component to interpreting observed effects (Wilkinson and APA Task Force, 1999). Cronbach and Shavelson (2004) suggested that researchers report a reliability coefficient for their specific study and not rely on published psychometrics, due to sampling and random errors. As with any point estimate, reliability coefficients are estimates of population parameters and tend to vary from sample to sample. This point is explicitly highlighted in reliability generalization studies that examine reliability fluctuation across studies (Vacha-Haase, Henson, & Caruso, 2002; Vacha-Haase & Thompson, 2011). Therefore, estimating the standard error of reliability coefficients with confidence intervals is critical.

A confidence interval (CI) provides information about the standard error of sample statistics and estimated range of values that most likely capture the true parameter (Cumming, 2012; Cumming & Finch, 2005). Larger standard errors and wider CIs are associated with low score reliability. The nominal width of a CI quantifies uncertainty and provides information regarding the precision of a point estimate (Cumming & Fidler, 2009). The standard error for the sample reliability coefficient is sensitive to sample size, the number of items, inter-item correlations, and homogeneity of variance (Duhachek & Iacobucci, 2004).

The sampling distribution for coefficient alpha follows a typical *F* distribution for large sample sizes (Feldt, 1965; Kristof, 1963). The sampling distribution of ordinal coefficient alpha can be conceived as having similar properties as coefficient alpha (B. Zumbo, personal communication, December 13, 2013). The reasoning is that ordinal alpha is simply coefficient alpha on the latent response distribution. The computation for ordinal alpha remains the same as that for coefficient alpha, except ordinal alpha is computed on the underlying latent continuous variable whereas coefficient alpha is computed on the measured continuous variable. From this point of view, the polychoric methods can be thought of as classes of complex transformations so that any property of coefficient alpha will likely hold for ordinal alpha. Therefore, the sampling distribution of ordinal alpha is highly likely to follow that of coefficient alpha. A natural question that follows is whether the methods for confidence intervals of coefficient alpha are appropriate to be used with ordinal alpha.

Consider the following four CI methods developed initially for coefficient alpha: Feldt (1965), Fisher (1950), Bonett (2002), and Hakstian and Whalen (1976). The differences between the methods are procedural transformations of sample coefficient alpha and the computation of variance. The formulae for the Feldt (1965) interval computations are

$$CI_{upper} = 1 - \left[\left(1 - \hat{\alpha}\right) \times F_{(\gamma/2), df_1, df_2} \right], \tag{6}$$

$$CI_{lower} = 1 - \left[\left(1 - \hat{\alpha}\right) \times F_{(1-\gamma/2), df_1, df_2} \right], \tag{7}$$

where $\hat{\alpha}$ is the sample reliability coefficient, γ is the specified level of significance, F represents the values at γ and $\gamma/2$ levels, n is the sample size with k items where $df_1 = (n - 1)$ and $df_2 = (n - 2)(k - 1)$.

Several methods transform sample coefficient alpha so confidence intervals may be computed within a normal theory framework. First, Fisher (1950) normalized a product moment correlation, such that

$$\text{Fisher's } z = \tfrac{1}{2} \ln\left(\frac{1 + |\hat{\alpha}|}{1 - |\hat{\alpha}|} \right), \tag{8}$$

where Fisher's z is the transformed parameter estimate and $\hat{\alpha}$ is the sample reliability coefficient. The z critical value (crit_z) is determined by the level of confidence where 95% has a value of 1.96. The standard error of z is estimated as

$$SE_z = \frac{1}{\sqrt{n-3}} \tag{9}$$

and applied to the lower and upper bounds of CI respectively as $\exp^{(2 \times (\text{Fisher}_z \pm \text{crit}_z \text{SE}_z)) - 1} / \exp^{(2 \times (\text{Fisher}_z \pm \text{crit}_z \text{SE}_z)) + 1}$. The transformed z statistic can be appropriately computed within a normal theory framework for a confidence interval and transformed back into the original units (Romano et al., 2011). Bonett (2002) extended Fisher's z (1950) as

$$\text{Bonett's } z = \ln\left(1 - |\hat{\alpha}|\right), \tag{10}$$

where Bonett's z statistic is closely normally distributed compared to that of sample coefficient alpha. The variances of Fisher's (1950) and Bonett (2002) z statistics are the same, yielding the lower and upper limits of CI respectively as $1 - \exp^{(\text{Bonett}_z \pm \text{crit}_z \text{SE}_z)}$. Hakstian and Whalen (1976) suggested another transformation of alpha, such that:

$$z = \left(1 - \hat{\alpha}\right)^{\frac{1}{3}}, \tag{11}$$

where the resulting z statistic is normally distributed with a variance, σ^2 of

$$\sigma^2 = \frac{18k(n-1)(1-\hat{\alpha})^{\frac{2}{3}}}{(k-1)(9n-11)^2}. \tag{12}$$

The lower and upper limits of CI respectively are $\Pr\left[1-c^{*3}(1-\hat{\alpha})^{\frac{1}{3}} \pm z_{1-(\alpha/2)}\sigma\right]^3$ where

$$c^{*3} = \frac{(9n-11)(k-1)}{9(n-1)(k-1)-2}. \tag{13}$$

Hakstian and Whalen (1976) argued their method is generally less biased than that of Fisher's (1950) z transformation, because they used the correction term $(1-\hat{\alpha})^{\frac{1}{3}}$ as an estimate of $(1-\alpha)^{\frac{1}{3}}$. There are notable performance differences among these CI methods as noted in the literature.

Recent Developments

Although several simulation studies have analyzed the performance of various methods for confidence intervals for coefficient alpha (Cui & Li, 2012; Iacobucci & Duhachek, 2003; Padilla et al., 2012; Romano et al., 2011), no known published study has analyzed confidence intervals for ordinal alpha. Romano et al. (2011) found negligible differences between the following eight confidence interval methods, with respect to bias, coverage, and precision for coefficient alpha computed for ordinal data: (a) Maydeu-Oliveres et al. (2007) asymptotic distribution free (ADF), (b) Bonett (2002), (c) Feldt (1965), (d) Fisher (1950), (e) Hakstian and Whalen (1976), (f) Duhachek and Iacobucci (2004), (g) Koning and Franses asymptotic (2003), and (h) Koning and Franses exact (2003) method. The findings suggest the ADF method was the least accurate for small sample sizes, and little was gained from departing from the Fisher approach. This finding is especially noteworthy because many other simulation studies suggested that ADF method outperformed other normal theory approaches, and that the Fisher approach yielded low coverage rates (Duhachek & Iacobucci, 2004; Hakstian & Whalen, 1976; Maydeu-Olivares et al., 2007). Romano et al. (2011) provided evidence that sophisticated CI methodology does not necessarily yield better performance. However, Romano et al. (2011) computed alpha coefficient for ordinal data based on Pearson and not polychoric correlations.

Romano et al.'s (2011) findings are important because advancements of ADF methods were considered the most robust in skewed distributions and small sample sizes. van Zyl et al. (2000) derived an asymptotic (i.e. large sample) distribution for sample coefficient alpha, only assuming a multivariate normal distribution and positive-definite matrix (Maydeu-Olivares, et al., 2007). Although van Zyl et al.'s (2000) intervals have been shown to yield the most narrow intervals, they often have undercoverage (Cui & Li, 2012).

Duhachek and Iacobucci (2004) extended van Zyl et al.'s (2000) method and presented statistics for coefficient alpha's standard error and computed an ADF-based CI. They found ADF intervals repeatedly outperformed other normal theory based intervals, including Feldt (1965) and Hakstian and Whalen (1976). This finding was consistent across all study conditions, but their study was not generalizable to Likert-type data. Maydeu-Olivares et al. (2007) found that the empirical coverage rate of the ADF intervals for coefficient alpha outperforms that of normal theory intervals, regardless of observed skewness and kurtosis of item distributions (Cui & Li, 2012; Romano et al., 2011). These results are significant because researchers are no longer bound by normality assumptions (i.e. normal theory) that were often violated when analyzing Likert-type data. Padilla et al. (2012) found that the normal theory bootstrap method had the most acceptable coverage rate followed by Bonett, and normal theory for non-normal data. Fisher method yielded unacceptably high variability, except when the scale had more than 15 items.

In sum, there is a need to evaluate the performance of the confidence intervals of ordinal alpha because Likert-type data is very commonly used in educational and behavioral research. Therefore, the present study investigated the coverage rates, widths, and biases of the four types of CIs (Bonnet, Feldt, Fisher, and Hakstian Whalen), and the accuracy of the ordinal alpha point estimates. Sample size, the number of items, skewness of the responses, and population alpha were varied.

Method

The program code was written using R (Version 3.0.2) using the R Studio interface (Version 0.98.976). The code was executed in a Windows-based environment (Version 8). Based on Maydeu-Olivares et al. (2007) and Hakstian and Whalen (1976) we generated the data from the factor analytic classical test theory model, assuming the parallel items model as follows:

a) For a given condition, generate a population of 1 million subjects by k number of items, with p population alphas, c response categories, and s skewness.

b) For a given sample size n, generate a $n \times k$ theoretical ability matrix θ^* such that $\theta^* \sim N(0,1)$.

c) Generate a $n \times k$ random error matrix U such that $U \sim MVN(0,\sigma)$ where

$$\sigma = \begin{bmatrix} .25 & 0 & \dots & 0 \\ 0 & .25 & \dots & \dots \\ \dots & \dots & .25 & 0 \\ 0 & \dots & 0 & .25 \end{bmatrix}.$$

d) Calculate the $n \times k$ matrix X^* such that $x_{ik} = \lambda\theta_{ik} + u_{ik}$ where λ values are specified below.

e) Categorize the scores in the item response distributions in X^* by applying rigid thresholds, τ (Muthén & Kaplan, 1985; Zumbo et al., 2007) to generate a skewed and symmetric distribution. The exact threshold values are provided in Table 1.

Design Factors

Population alphas (a). Three population alphas were specified at .6, .8, and .9 as used in previous simulation studies (Cui & Li, 2012; Padilla et al., 2012; Romano et al., 2011; Zumbo et al., 2007). Factor pattern coefficients values (λs) were based on Zumbo et al. (2007) with values of .311, .471, and .625 for population ordinal coefficient alphas of .6, .8, and .9, respectively.

Sample size (n). The design conditions included four levels of sample size (20, 50, 100, 200). The sample sizes were selected based on previous studies and represent sample sizes often noted in applied research (Cui & Li, 2012; Duhachek & Iacobucci, 2004; Maydeu-Olivares et al., 2007; Padilla et al., 2012; Romano et al., 2011; Yuan & Bentler, 2002; Zumbo et al., 2007). While large sample sizes are always desirable, they are not always realistic. Duhachek and Iacobucci (2004) indicated that sample sizes beyond 200 have diminishing returns for coefficient alpha, given a sufficient number of items and strong inter-item correlations. Therefore, they were not simulated in the present study. Similarly, we considered sample sizes as small as 20 because this is not an uncommon sample size in educational research and has therefore been included as a data condition in other similar simulation studies (e.g. Natesan & Thompson, 2007). Moreover, considering a sample size as low as 20 helps the researcher understand a possible lower bound of sample size necessary for estimating ordinal alpha.

Number of items (k). The number of items chosen were $k = 5$, 10, 25,and 40. Previous studies have simulated between two and 40 items, which also reflects the test widths of interest to applied researchers (Cui & Li, 2012; Duhachek & Iacobucci, 2004; Maydeu-Olivares et al., 2007; Padilla et al., 2012; Romano et al., 2011; Yuan & Bentler, 2002; Zumbo et al., 2007). It is not uncommon to consider 5 and 10 Likert-type items per factor in simulation studies (e.g. Ankemann & Stone, 1992; Kieftenbeld & Natesan, 2012; Reise & Yu, 1990). Forty items were considered as the upper bound of test length.

Skewness (s). Two types of observed item response distributions were selected: $s = 0, -1.217$. These values were selected to demonstrate the impact of symmetry on precision of confidence intervals for ordinal coefficient alpha (Zumbo et al., 2007). Threshold values are used to categorically score the individual item's value computed in steps (a) through (e) described above. The following thresholds for the two item response distributions and relative response categories are based on the works of Zumbo et al. (2007) and specified in Table 1.

Response categories (C). Two scales of response categories (C) were selected: the five-point and seven-point scales. Duhachek and Iacobucci (2004) demonstrated that confidence interval performance does not improve beyond seven response categories. Therefore, simulating more than seven response categories was not deemed necessary. A five-point Likert scale is commonly used in behavioral research. The resulting design is a fully-crossed $2(s) \times 2(C) \times 3(\alpha) \times 4(n) \times 4(k)$ factorial design with 192 conditions.

Table 1. Likert Scale Thresholds

Skewness	Five-point Scale		Seven-point Scale	
y_j	0	-1.217	0	-1.217
1	$y_j^* \leq -1.8$	$y_j^* > 1.8$	$y_j^* \leq -2.14$	$y_j^* > 2.4$
2	$-1.8 < y_j^* \leq -0.6$	$1.8 \geq y_j^* > 1.34$	$-2.14 < y_j^* \leq -1.29$	$2.4 \geq y_j^* > 1.95$
3	$-0.6 < y_j^* \leq 0.6$	$1.34 \geq y_j^* > 0.77$	$-1.29 < y_j^* \leq -0.43$	$1.95 \geq y_j^* > 1.42$
4	$40.6 < y_j^* \leq 1.8$	$0.77 \geq y_j^* > 0.05$	$-0.43 < y_j^* \leq 0.43$	$1.42 \geq y_j^* > 0.99$
5	$51.8 < y_j^*$	$0.05 \geq y_j^*$	$-0.43 < y_j^* \leq 0.43$	$0.99 \geq y_j^* > 0.47$
6	NA	NA	$0.43 < y_j^* \leq 2.14$	$0.47 \geq y_j^* > -0.2$
7	NA	NA	$2.14 < y_j^*$	$-0.2 \geq y_j^*$

Diagnostics

Coverage rates were computed as,

$$\text{coverage rate} = (A/B) \times 100\%, \tag{14}$$

where A is the frequency of intervals which contain the true population parameter, and B is the total number of intervals. Coverage rate should have a value close to the nominal level but are not a sufficient diagnostic, particularly when the skewness of the sampling distribution is not provided or unknown (Jennings, 1987; Schall, 2012; Zhang, Gutiérrez Rojas, & Cuervo, 2010). In addition to coverage rate, positive and negative bias of intervals that do not contain the true value must be reported. Among the $(B - A)$ intervals that did not contain the true value, the number of intervals which were below and above the true value when expressed as percentage of the total number of intervals indicate negative and positive bias of CI, respectively. An imbalance in these biases indicate possible systematic bias in the estimation (Natesan, 2015). An unbiased interval is equally likely to be above or below the true value. Therefore an unbiased CI estimate would have roughly equal number of negatively and positively biased intervals. CI width is the difference between the upper and lower limits of the CI. Following Zhang et al. (2010), mean and variance of CI widths were computed. A highly variable CI width indicates poor precision for the interval estimate method.

Precision of point estimates. Bias and RMSE of ordinal alpha were computed. Bias is the difference between the true population parameter value and the sample estimate. RMSE across N replications is computed as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(x_i - \mu\right)^2}{N}}, \tag{15}$$

where μ is the true population parameter value, x_i is the estimate of ordinal alpha in the i^{th} replication. To determine whether a confidence interval is unacceptably wide, the empirical standard error (SE) of ordinal alpha was computed. The empirical SE is the standard deviation of all sample ordinal alpha estimates for a given condition.

Data Analysis

Following the simulation, η^2 effect sizes were examined for separate ANOVAs to understand the variance in the simulation diagnostics explained by the data conditions. The independent variables were population alphas, sample size, number of items, skewness, and response categories. The dependent variables were coverage rates, CI width, variance of CI width, RMSE, and bias of ordinal alpha. Both main effects and all higher order interactions were examined. Following Cohen (1988), 1%, 6%, and 14% were considered small, medium, and large effect sizes for η^2. Only large main and higher order interaction effects are interpreted and discussed.

Results

Practical Computation Issues

To minimize the standard error of the simulation, 1,000 samples were drawn for each condition (Fan & SAS Institute, 2002; Wang & Thompson, 2007). The Appendix includes a list of 25 conditions (of the 192 conditions) that did not execute due to repeated crashing. The error message stated that a “not positive-definite matrix” was computed which caused computations to stop. This error occurred when there was a large number of items (e.g. 25 or 40 items) with a small sample size (e.g. $n = 20$). The rigid thresholds set for the 5-point and 7-point Likert scales removed important differences between the available response options (1, 2, … 5 or 1, 2, … .7). Ultimately, there simply was not enough sampling variability generated across each repetition and the variables became constant (e.g. all responses were scored “3”). When no variability was generated either across items or subjects, the covariance and standard deviation are essentially zero. When this occurs, estimation stops because one cannot divide by a standard deviation of zero to compute polychoric correlation. The issues related to the lack of variance generated seem to be an artifact of restricted range with the ordinal data. The resulting dataset contained 167,000 replications (167 workable conditions × 1,000 samples each). The total time elapsed was approximately 691 computing hours. The simulations were executed on a Dell Precision T3600 Intel (R) Xeon (R) CPU E5-1620 3.60 GHz Windows 8 machine.

Coverage rate and CI bias.

Overall, coverage rates were much lower than the 95% nominal rate as seen in Figure 1, ranging from 46% (Feldt) to 62% (Fisher). Feldt method had the lowest coverage rate due to the confounding impact of several independent variables summarized in Table 2. ANOVA results show that skewness explained most of the variance in mean coverage rates (23.651% to 62.915%) except for the Feldt method. Interaction effects have a dominating presence, especially for the Feldt method where η^2 = 64.968%.

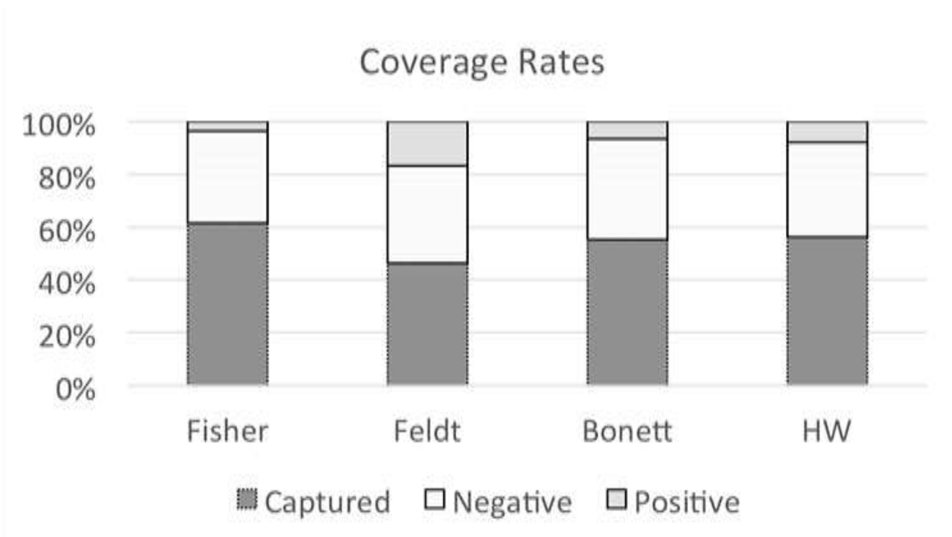

Figure 1. Coverage rates by confidence interval method across N = 167,000 replications

Table 2. η^2 (%) by confidence interval method for coverage rates

Independent variable	*Fisher*	*Feldt*	*Bonett*	*HW*
population alpha (α)	0.422	0.252	0.110	0.381
sample size (n)	4.122	7.453	4.692	7.085
items (k)	1.934	2.722	5.312	6.116
skewness (s)	62.915	23.651	55.108	46.769
response categories (C)	2.157	0.322	0.997	1.034
Interactions[a]	27.197	64.968	32.756	37.590
$\alpha \times s$	9.375	13.129	13.165	16.919
$\alpha \times n$	4.764	6.705	6.307	5.371
$\alpha \times n \times k$		6.430		
$\alpha \times n \times k \times C$		8.979		

Note. [a] Interactions includes all possible 2nd, 3rd, 4th, and 5th order interactions; only η^2 (%) values greater than 4% are reported for the interaction terms

The largest two-way interaction effect, population alpha by skewness, is shown in Figure 2. Two-way mean interaction plots of population alpha by skewness effects. As population alpha increased from .6 to .9, the estimated marginal means (EMMs) of captured coverage rates increased for skewed distributions. Coverage rates increased due to the joint influence of population alpha levels and skewness. Coverage rates were higher in skewed data except for the Feldt method, when population alpha levels were .6. The CIs that did not contain the true value more often underestimated ordinal alpha for all methods. That is, negatively biased intervals (35-38%) occurred more frequently than positively biased intervals (4-17%) as shown in Figure 2.

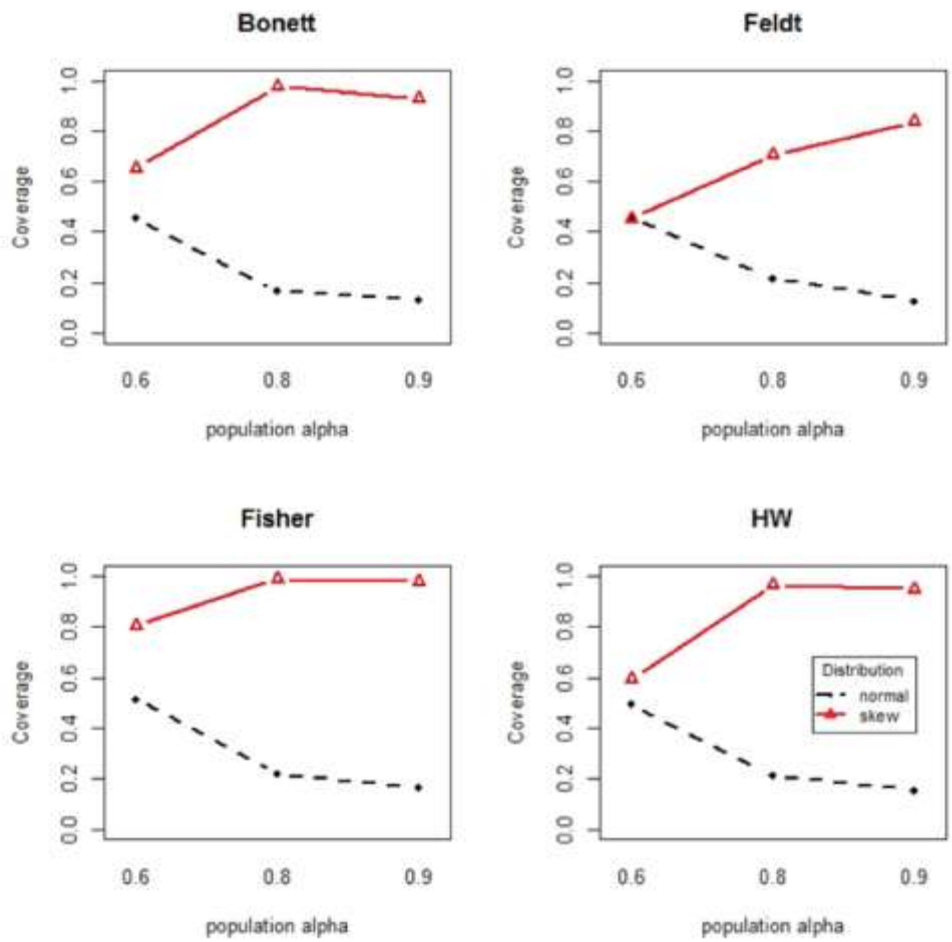

Figure 2. Two-way mean interaction plots of population alpha by skewness effects on coverage rates.

CI widths. The boxplot shown in Figure 3 depicts the interquartile ranges of the 95% CI width for the four confidence interval methods. The Fisher confidence intervals consistently yielded the narrowest intervals, while Bonett intervals were the widest across all conditions. Table 3 shows that all intervals became narrower with increase in sample size and population alpha with one exception. The exception occured with the Feldt interval when n = 200 and the population alpha increased from .8 to .9.

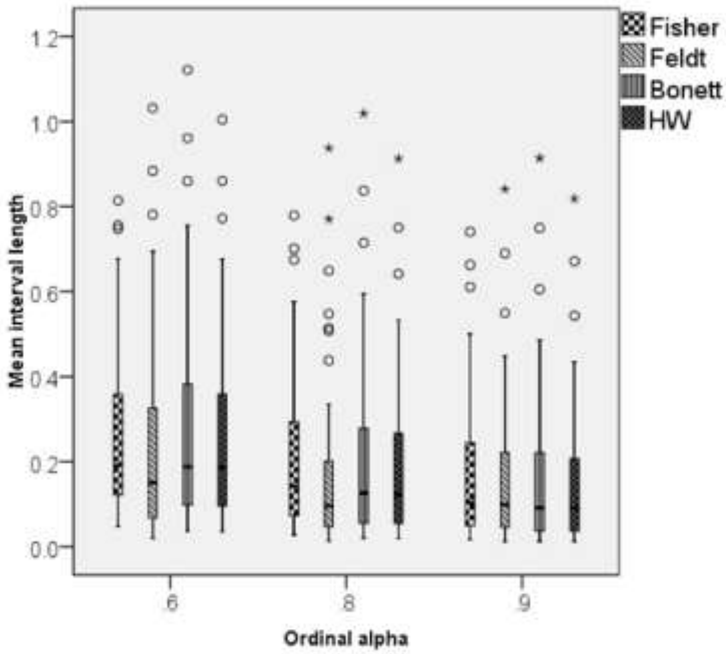

Figure 3. Boxplot of 95% CI widths for four estimation methods

Table 3. Mean confidence interval width (SD) at study condition level

Method	Population alpha	Sample Size 20	50	100	200
Fisher	0.6	.654 (.073)	.286 (.032)	.184 (.017)	.133 (.010)
	0.8	.573 (.082)	.223 (.029)	.140 (.015)	.091 (.008)
	0.9	.510 (.077)	.184 (.024)	.108 (.014)	.060 (.007)
Feldt	0.6	.673 (.125)	.254 (.041)	.153 (.019)	.100 (.001)
	0.8	.524 (.114)	.178 (.029)	.103 (.014)	.071 (.008)
	0.9	.499 (.107)	.162 (.026)	.094 (.014)	.073 (.008)
Bonett	0.6	.742 (.140)	.281 (.045)	.172 (.022)	.126 (.013)
	0.8	.610 (.135)	.202 (.033)	.126 (.017)	.080 (.009)
	0.9	.516 (.112)	.165 (.027)	.093 (.013)	.050 (.006)
HW	0.6	.665 (.125)	.270 (.043)	.168 (.021)	.125 (.013)
	0.8	.547 (.121)	.194 (.031)	.123 (.017)	.079 (.009)
	0.9	.463 (.100)	.158 (.026)	.091 (.013)	.050 (.006)

Shown in Figures 4 through 6, the CIs became narrower with increase in both sample size and the number of items simultaneously for all methods. The intervals became quite narrow when sample size = 200 and the number of items = 40. There are no striking visual differences in the confidence interval widths between the Fisher, Feldt, Bonett, and HW methods across various levels of population alpha because the patterns are similar for all methods.

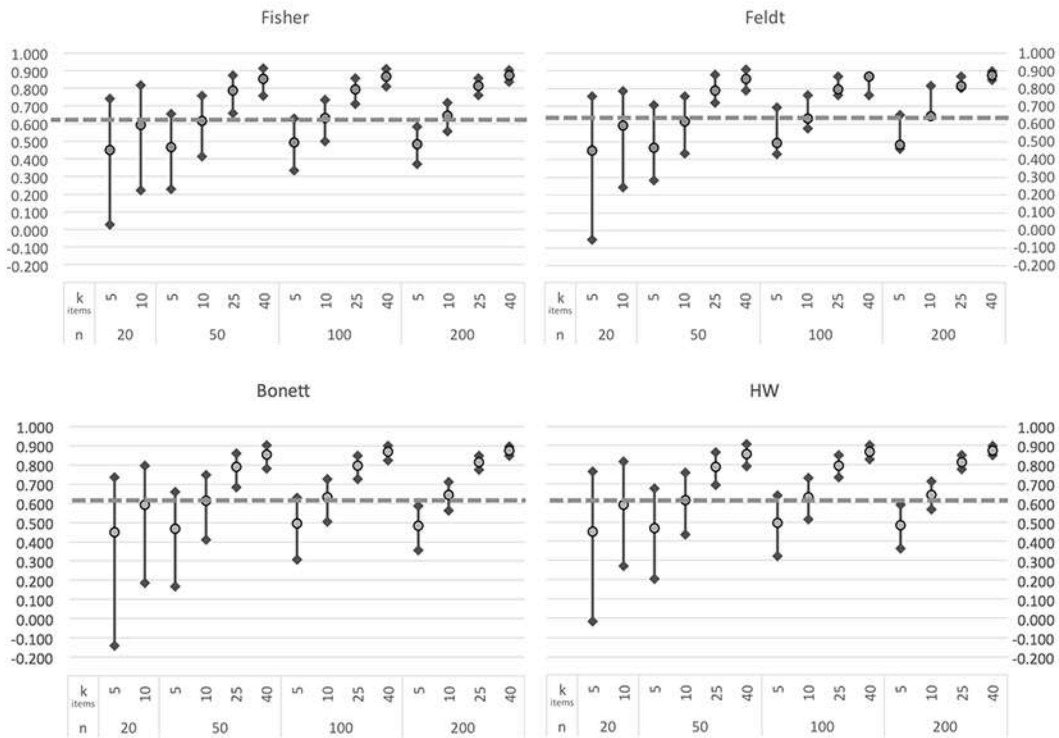

Figure 4. Mean confidence limits for population alpha = .6. Dashed line references population parameter. Bottom marker is the mean lower limit and top marker is the mean upper limit. Middle marker is the mean sample ordinal coefficient alpha.

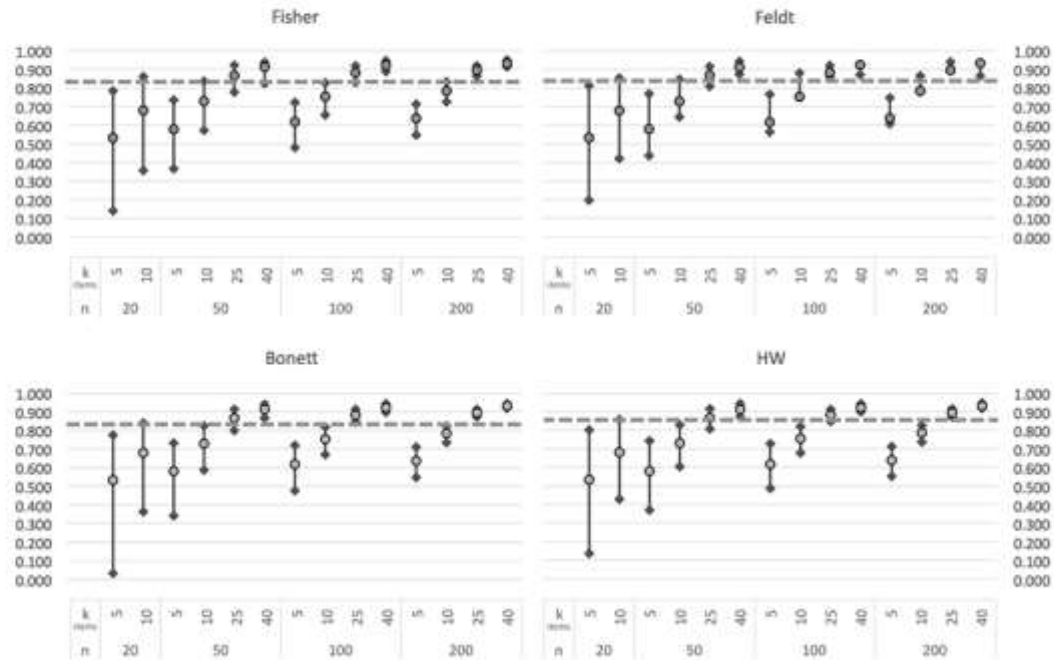

Figure 5. Mean confidence limits for population alpha = .8. Dashed line references population parameter. Bottom marker is the mean lower limit and top marker is the mean upper limit. Middle marker is the mean sample ordinal coefficient alpha.

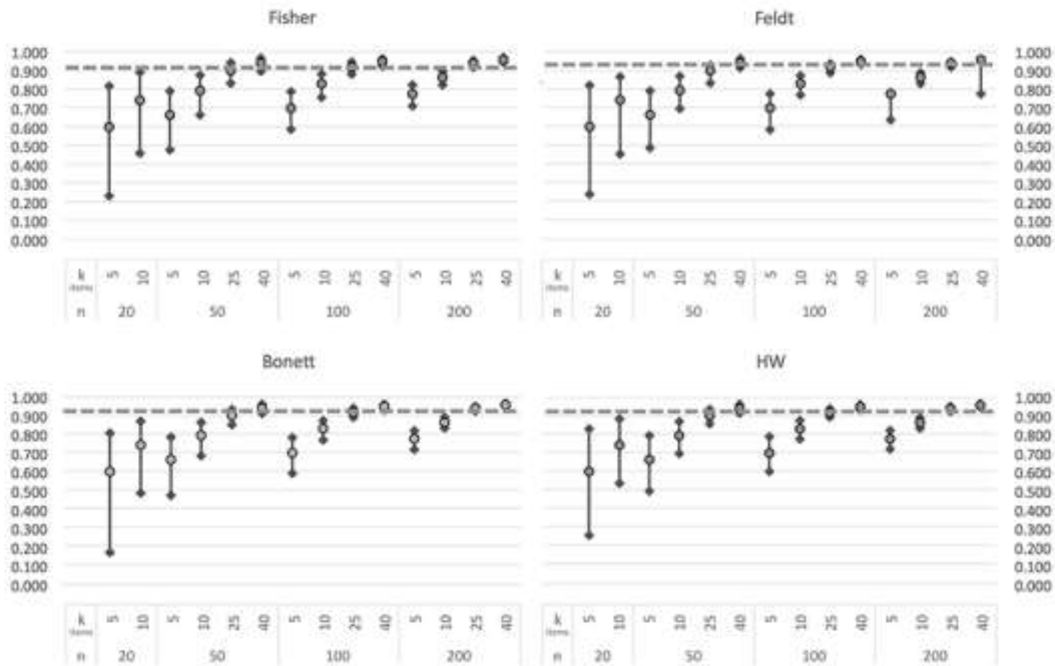

Figure 6. Mean confidence limits for population alpha = .9. Dashed line references population parameter. Bottom marker is the mean lower limit and top marker is the mean upper limit. Middle marker is the mean sample ordinal coefficient alpha.

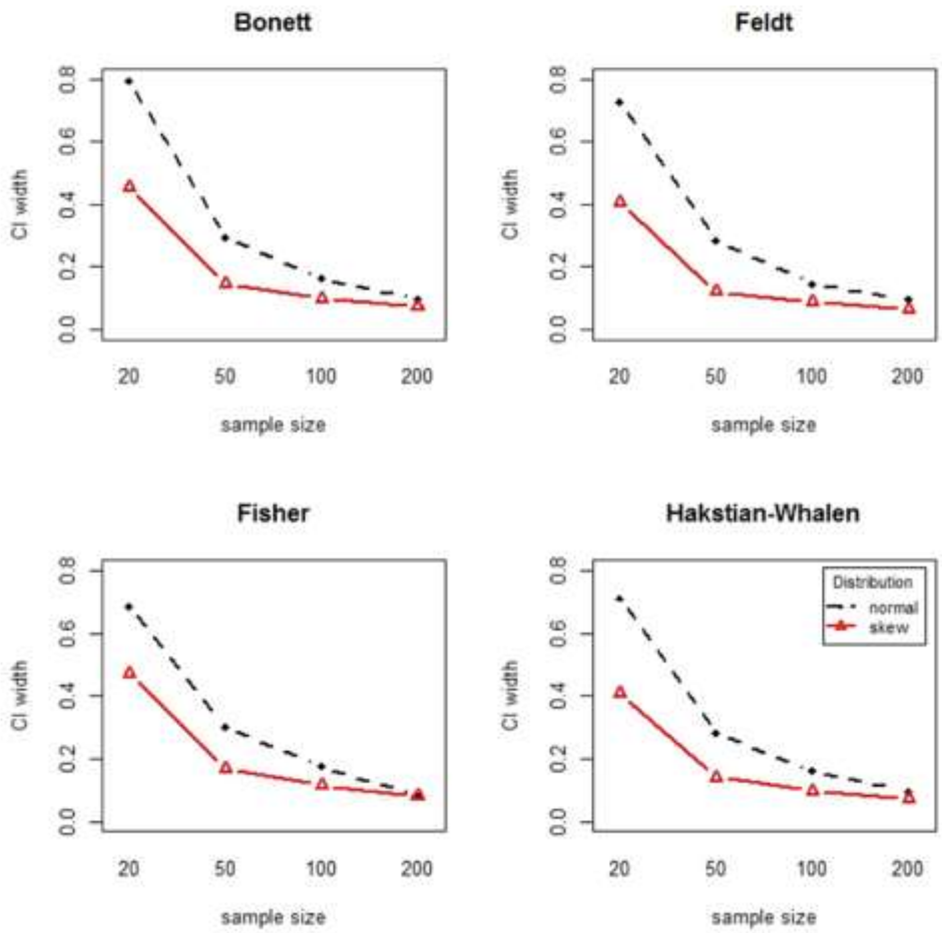

Figure 7. Interaction of sample size by skewness on CI width

As seen in Table 4, sample size explained the most variance in CI width across all methods. Specifically, the CI widths decreased with increase in sample size. Of the interaction effects, the largest amount of the interaction effects were explained by the sample size by skewness interaction. The sample size by skewness interactions ranged from 2.55% (Fisher) to 5.03% (Bonett). While the η^2 values for both interactions may be considered "small," (i.e., < 1%, Cohen, 1988), the implications are meaningful. The CI widths were consistently smaller across all methods when sample size increased to 200 and skewness = 0 as shown in Figure 7.

Table 4. η^2 (%) by confidence interval method for CI width

Independent variable	*Fisher*	*Feldt*	*Bonett*	*HW*
population alpha (α)	3.882	2.516	3.852	4.290
sample size (n)	67.254	55.737	57.940	55.062
items (k)	13.599	11.795	14.709	16.566
skewness (s)	6.053	7.914	6.527	7.068
Interactions[a]	5.280	15.487	10.148	10.010

Note. [a] Interactions includes all possible interactions. Response categories had η^2 less than 1%

Variance of CI widths. Effect sizes summaries for variance CI widths for all CI methods are shown in Table 5. Sample size by number of items had the largest effect across the four confidence interval methods. The η^2 values ranged from 2.140% (Fisher) to 10.394% (HW) with the mean plots provided below in Figure 8. All four methods followed the same pattern with variance of CI widths sharply decreasing as both sample size and the number of items increased. In summary, the joint influence of the number of items and sample size impacted the mean variance of CI width across all methods.

Table 5. η^2 (%) by confidence interval method for variance of CI width

Independent variable	*Fisher*	*Feldt*	*Bonett*	*HW*
population alpha (α)	0.144	0.500	0.738	0.829
sample size (n)	83.358	63.538	67.264	66.409
items (k)	2.821	5.989	7.285	7.915
skewness (s)	0.254	3.011	2.257	2.492
Interactions[a]	12.780	26.866	22.378	22.276

Note. [a] Interactions includes all possible 2nd, 3rd, 4th, and 5th order interactions; number of categories had $\eta^2 < 1\%$

Point estimates. Given the large η^2 values of sample size (43.377%) and items (34.102%), further post-hoc analyses were conducted to see which levels of the independent contributed the most to the variance of RMSE as seen in Table 6. The RMSEs decreased from .083 to .022 as sample size increased from 20 to 200. As the number of items increased from 5 to 40, the RMSE values decreased from .068 to .011.

Overall, sample ordinal alpha was negatively biased (M = −.054, SD = .103, N = 167,000) ranging from −.69 to .328. The distribution of bias was negatively skewed (−1.10, SE = .006) with a leptokurtic shape (1.940, SE = .012). Skewness levels (0, −1.217) had the largest impact on the bias of ordinal coefficient alpha. Sample size × skewness explained 5.36% of the variance in bias. The rest of the interactions explained less than 4% of the variance. Negative skewness resulted in a less biased estimate (EMM = .013, SE < .001) compared to no skewness (EMM = −.122, SE < .001). These results support the use of ordinal coefficient alpha when analyzing Likert-type or ordinal data because less bias is present when data are skewed. In summary, the precision of ordinal coefficient alpha, in terms of RMSE, is best explained by the main effects of sample size and the number of items. Bias is best explained by the main effect of skewness and a combination of small interaction effects.

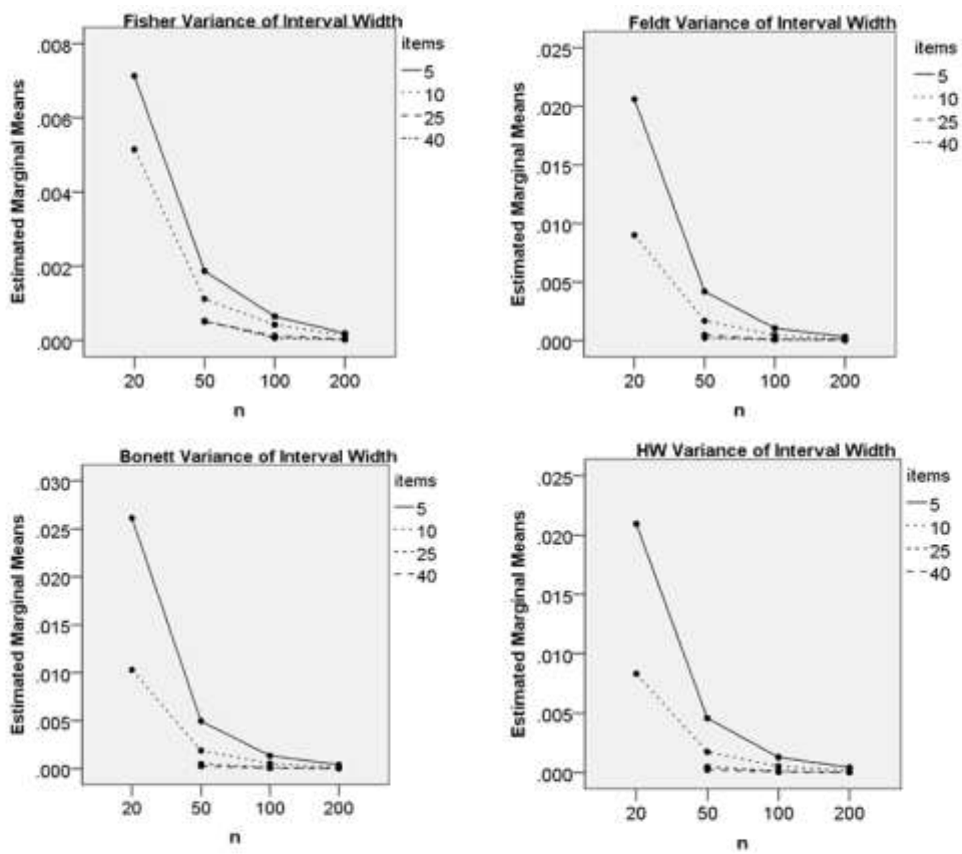

Figure 8. Interaction of samples size by number of items on CI width

Table 6. η^2 (%)for RMSE and bias

Source	*RMSE*	*Bias*
population alpha (α)	5.295	2.811
sample size (n)	42.377	10.376
items (k)	34.102	1.600
skewness (s)	10.094	41.754
response categories (C)	0.036	2.166
Interactions[a]	8.084	19.085

Conclusion

The aim in this study was to evaluate the performance of Feldt, Fisher, Bonett, and HW confidence interval methods for ordinal coefficient alpha. The simulation findings are only applicable to study designs where the sample sizes range from 20 to 200, the number of items range from five to 40, scores are categorized into symmetric and skewed item response distributions, with five or seven response categories. None of the CI methods suggested for coefficient alpha have adequate coverage for ordinal alpha. Skewness had the largest impact on coverage rates. Mean coverage rates were 46% - 62%, low, and unacceptably low for all methods. This could lead to high type-I error rates. Moreover, for all methods the CIs that did not contain the true values were more negatively biased. Clearly these findings show the need for a new method that specifically formulates CI for ordinal alpha.

CI widths were statistically significantly different across Feldt, Fisher, Bonett, and HW methods ($p < .05$). CI widths became narrower as population alphas increased and sample size increased. There are small, but notable differences observed with CI width between methods. The Feldt method is the only CI method that did not use any transformation of sample ordinal coefficient alpha, and was therefore, impacted differently than Fisher, Bonett, and HW. The Feldt CI width is determined as a function of the degrees of freedom based on sample size and number of items, and was therefore heavily impacted by the interactions of these conditions. The Fisher, Bonett, and HW methods apply logarithmic transformations of sample ordinal alpha and were therefore more easily explained by varying sample size and the number of items.

Sample size and number of items best explained the precision of ordinal alpha. Interestingly, the number of response categories is a strong predictor of coefficient alpha, but not necessarily for ordinal coefficient alpha (Zumbo et al., 2007). EMMs of RMSE were statistically significant across all levels of sample size and number of items. The practical implications suggest keeping an instrument, with Likert-type data, between 10-25 items, while striving for at least 50 participants. However, researchers should not use any of the CI formulae tested in the present study to compute confidence interval for ordinal alpha.

Bias was best explained by skewness, and sample size by skewness interaction effect. Overall, bias is persistently negative across all design levels except for skewed data. Bias approached zero when $n = 200$. This shows that regardless of the method, when the estimate is more biased the coverage rate will be lower. Again, very little confidence should be placed on confidence intervals methods for ordinal alpha.

As with any simulation study, the results are limited to the conditions specified. The conditions were justified with previous research to portray scenarios in applied research. The conclusions hold for the study conditions specified; therefore, a number of opportunities exist to extend the current research. First, a confidence interval method specifically for ordinal alpha which improves coverage rates closer to the nominal rate needs to be developed. Additionally, the contiguous points between 10 and 25 items may be explored to determine the optimal point of precision of ordinal alpha for both RMSE and bias.

References

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

Allen, M. J., & Yen, W. M. (2002). *Classical true-score theory*. Prospect Heights, IL: Waveland Press.

Ankemann, R. D., & Stone, C. A. (1992, April). *A Monte Carlo study of marginal maximum likelihood parameter estimates for the graded response model*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics, 27*(4), 335-340. doi: 10.3102/10769986027004335

Cohen, J. (1994). The earth is round (p<.05). *American Psychologist, 49*(12), 997-1003. doi: 10.1037/0003-066x.49.12.997

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Crocker, L., & Algina, J. (2008). *Introduction to classical & modern test theory*. Mason, OH: Cengage Learning.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334. doi: 10.1007/bf02310555

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*(3), 391-418. doi: 10.1177/0013164404266386

Cui, Y., & Li, J. (2012). Evaluating the performance of different procedures for constructing confidence intervals for coefficient alpha: A simulation study. *British Journal of Mathematical and Statistical Psychology, 65*(3), 467-498. doi: 10.1111/j.2044-8317.2012.02038.x

Cumming, G. (2012). *Understanding the new statistics*. New York: NY: Taylor & Francis Group.

Cumming, G, & Fidler, F. (2009). Confidence intervals: Better answers to better questions. *Zeitschrift für Psychologie/Journal of Psychology*, *217*(1), 15-26. doi: 10.1027/0044-3409.217.1.15

Cumming, G, & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist*, *60*(2), 170-180. doi: 10.1037/0003-066X.60.2.170

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. doi: 10.1037/0021-9010.89.5.792

Fan, X, & SAS Institute. (2002). SAS® for Monte Carlo studies: A guide for quantitative researchers. Cary, NC: SAS Institute.

Fan, X, & Thompson, B. (2001). Confidence intervals for effect sizes: Confidence intervals about score reliability coefficients, please: An EPM guidelines editorial. *Educational and Psychological Measurement*, *61*(4), 517-531. doi: 10.1177/0013164401614001

Feldt, L. S. (1965). The approximate sampling distribution of Kuder-Richardson coefficient twenty. *Psychometrika, 30*(3), 357-370. 10.1007/bf02289499

Finch, S., Cumming, G., & Thomason, N. (2001). Colloquium on effect Sizes: The roles of editors, textbook authors, and the publication manual: Reporting of statistical inference in the Journal of Applied Psychology: Little evidence of reform. *Educational and Psychological Measurement, 61*(2), 181-210. doi: 10.1177/0013164401612001

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

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*(4), 466-491. doi: 10.1037/1082-989x.9.4.466

Gadermann, A. M., Guhn, M, & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research & Evaluation*, *17*(3). Available online: http://pareonline.net/getvn.asp?v=17&n=3

Hakstian, A. R., & Whalen, T. E. (1976). A k-sample significance test for independent alpha coefficients. *Psychometrika, 41*(2), 219-231. doi: 10.1007/bf02291840

Iacobucci, D., & Duhachek, A. (2003). Advancing alpha: Measuring reliability with confidence. *Journal of Consumer Psychology, 13*(4), 478-487. doi: 10.1207/s15327663jcp1304_14

Jennings, D. E. (1987). How do we judge confidence-interval adequacy? *The American Statistician, 41*(4), 335-337. doi: 10.2307/2684768

Jöreskog, K.G., 1990. New developments in LISREL: Analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality & Quantity, 24*(4), 387-404. doi: 10.1007/bf00152012

Kieftenbeld, V. & Natesan, P. (2012). Recovery of graded response model parameters: A comparison of marginal maximum likelihood and Markov chain Monte Carlo estimation. *Applied Psychological Measurement*, *36*(5), 399-419. doi: 10.1177/0146621612446170

Koning, A. J., & Franses, P. H. (2003). *Confidence intervals for Cronbach's coefficient alpha values*. (No. ERS-2003-041-MKT). *ERIM Report Series Research In Management.* Retrieved from http://hdl.handle.net/1765/431

Kristof, W. (1963). The statistical theory of stepped-up reliability coefficients when a test has been divided into several equivalent parts. *Psychometrika, 28*(3), 221-238. doi: 10.1007/bf02289571

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. doi: 10.1037/1082-989x.12.2.157

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of*

Mathematical and Statistical Psychology, 38(2), 171-189. doi: 10.1111/j.2044-8317.1985.tb00832.x

Natesan, P. (2015). Comparing interval estimates for small sample ordinal CFA models. *Frontiers in Psychology*, *6*, 1599. doi: 10.3389/fpsyg.2015.01599

Natesan, P. & Thompson, B. (2007). Extending improvement-over-chance I-index effect size simulation studies to cover some small-sample cases. *Educational and Psychological Measurement, 67*(1), 59-72. 10.1177/0013164406292028

Novick, M. R., & Lewis, M. R. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika, 32*(1), 1-13. doi: 10.1007/bf02289400

Padilla, M. A., Divers, J., & Newton, M. (2012). Coefficient alpha bootstrap confidence interval under nonnormality. *Applied Psychological Measurement, 36*(5), 331-348. doi: 10.1177/0146621612445470

Reise, S. P. & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27*(2), 133-144. doi: 10.1111/j.1745-3984.1990.tb00738.x

Romano, J. L., Kromrey, J. D. , Owens, C. M. , & Scott, H. M. (2011). Confidence interval methods for coefficient alpha on the basis of discrete, ordinal response items: Which one, if any, is the best? *Journal of Experimental Education, 79*(4), 382-403. doi: 10.1080/00220973.2010.510859

Schall, R. (2012). The empirical coverage of confidence intervals: Point estimates and confidence intervals for confidence levels. *Biometrical Journal, 54*(4), 537-551. doi: 10.1002/bimj.201100134

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*(4), 350-353. doi: 10.1037/1040-3590.8.4.350

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 107-120. doi: 10.1007/s11336-008-9101-0

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80*(1), 99-103. doi: 10.1207/s15327752jpa8001_18

Thompson, B. (2006a). *Foundations of behavioral statistics.* New York, NY: The Guilford Press.

Thompson, B. (2006b). Research synthesis: Effect sizes. In J. L. Green, G. Camilli & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 583-603). New Jersey: Lawrence Erlbaum Associates.

Vacha-Haase, T., Henson, R. K., & Caruso, J. C. (2002). Reliability generalization: Moving toward improved understanding and use of score reliability. *Educational and Psychological Measurement, 62*(4), 562-569. doi: 10.1177/0013164402062004002

Vacha-Haase, T., & Thompson, B. (2011). Score reliability: A retrospective look back at 12 years of reliability generalization studies. *Measurement and Evaluation in Counseling and Development, 44*(3), 159-168. doi: 10.1177/0748175611409845

van Zyl, J. M., Neudecker, H., & Nel, D. G. (2000). On the distribution of the maximum likelihood estimator of Cronbach's alpha. *Psychometrika, 65*(3), 271-280. doi: 10.1007/bf02296146

Wang, Z., & Thompson, B. (2007). Is the Pearson r2 biased, and if so, what is the best correction formula? *Journal of Experimental Education, 75*(2), 109-125. doi: 10.3200/jexe.75.2.109-125

Wilkinson, L , & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594-604. doi: 10.1037/0003-066x.54.8.594

Yuan, K., & Bentler, P. (2002). On robustness of the normal-theory based asymptotic distributions of three reliability coefficient estimates. *Psychometrika, 67*(2), 251-259. doi: 10.1007/bf02294845

Zhang, H., Gutiérrez Rojas, H. A., & Cuervo, E. C. (2010). Confidence and credibility intervals for the difference of two proportions. *Revista Colombiana De Estadística, 33*(1), 63-88.

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods 6*, 21-29. doi: 10.22237/jmasm/1177992180

Appendix A: Non-Executable Conditions

Obs num	*SimID*	*Population ordinal coeff alpha*	*Factor loading*	*k items*	*c response categories*	*n sample size*	*Skewness*
1	3	0.9	0.625	25	5	20	0.000
2	4	0.6	0.311	40	5	20	0.000
3	7	0.6	0.311	25	7	20	0.000
4	8	0.8	0.471	40	7	20	0.000
5	12	0.9	0.625	40	5	50	0.000
6	35	0.8	0.471	25	5	20	-1.217
7	36	0.9	0.625	40	5	20	-1.217
8	39	0.9	0.625	25	7	20	-1.217
9	40	0.6	0.311	40	7	20	-1.217
10	67	0.6	0.311	25	5	20	0.000
11	68	0.8	0.471	40	5	20	0.000
12	71	0.8	0.471	25	7	20	0.000
13	72	0.9	0.625	40	7	20	0.000
14	99	0.9	0.625	25	5	20	0.000
15	100	0.6	0.311	40	5	20	-1.217
16	103	0.6	0.311	25	7	20	-1.217
17	104	0.8	0.471	40	7	20	-1.217
18	131	0.8	0.471	25	5	20	0.000
19	132	0.9	0.625	40	5	20	0.000
20	135	0.9	0.625	25	7	20	0.000
21	136	0.6	0.311	40	7	20	0.000
22	163	0.6	0.311	25	5	20	0.000
23	164	0.8	0.471	40	5	20	-1.217
24	167	0.8	0.471	25	7	20	-1.217
25	168	0.9	0.625	40	7	20	-1.217